2007

Determination of the Fundamental Image Categories for Typical Consumer Imagery

Kenneth N. Fleisher

Determination of the Fundamental Image Categories for Typical Consumer Imagery

Kenneth N. Fleisher
B.F.A. Rochester Institute of Technology
(1989)

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Color Science
in the Chester F. Carlson Center of Imaging Science of the College of Science
Rochester Institute of Technology

September 2007

Kenneth N. Fleisher

Signature of the Author

Roy S. Berns

Accepted by Dr. Roy S. Berns
Coordinator, M.S. Degree Program

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE
COLLEGE OF SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

M.S. DEGREE THESIS

The M.S. Degree Thesis of Kenneth N. Fleisher
has been examined and approved by two members of the color science faculty
as satisfactory of the thesis requirement for the Master of Science Degree

Mark D. Fairchild
Dr. Mark D. Fairchild

Roy S. Berns
Dr. Roy S. Berns

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE
COLLEGE OF SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

THESIS REPRODUCTION PERMISSION STATEMENT

Title of thesis: **Determination of the Fundamental Image Categories for Typical Consumer Imagery**

I, Kenneth N. Fleisher, hereby **grant permission** to the Wallace Library of the Rochester Institute of Technology to reproduce my thesis in whole or in part. Any reproduction will not be for commercial use or profit.

Kenneth N. Fleisher
Signature of the Author

Date

Determination of the Fundamental Image Categories for Typical Consumer Imagery

Kenneth N. Fleisher

B.F.A. Rochester Institute of Technology

(1989)

ACKNOWLEDGEMENT

I would like to extend my sincerest thank you to Dr. Ethan D. Montag for his guidance on this project and for helping me to learn a great deal about psychophysics.

I would also like to thank Dr. Roy S. Berns, Dr. Mark D. Fairchild , and everyone at the Munsell Color Science Laboratory for helping to make my journey in color science a truly memorable experience.

A special thanks to Dave Wyble and Lawrence Taplin for their help with the Distributed Experiment and to Li, Willy, Iris, Mahnaz, and Mahdi for listening to and helping me with an endless amount of random questions about color science.

This research was supported by a grant from Lexmark International, Inc. I want to thank Ann McCarthy and Brant Nystrom from Lexmark for making this opportunity possible.

And last, but far from least, I want to thank my parents for their support and encouragement throughout the past few years.

Determination of the Fundamental Image Categories for Typical Consumer Imagery

Kenneth N. Fleisher

B.F.A. Rochester Institute of Technology

(1989)

A thesis submitted in partial fulfillment of the requirements

for the degree of Master of Science in Color Science

in the Chester F. Carlson Center of Imaging Science of the College of Science

Rochester Institute of Technology

ABSTRACT

Many tasks in imaging science are image-dependent. While a particular dependency might simply be a function of certain physical attributes of an image, often it is closely related to the perceived semantic category. Therefore, a thorough understanding of image semantics would be of substantial practical value. The primary goal of this research was to determine the fundamental semantic categories for typical consumer imagery. Two psychophysical experiments were performed. Experiment I was a Free Sorting Experiment where observers were asked to sort 321 images into piles of similar images. Experiment II was a Distributed Experiment conducted over the internet which used the method of triads to collect similarity and dissimilarity data from 321 images. Due to the large number of images included in the experiment, the method of non-repeating random paths was employed to reduce the number of required responses. Both experiments were analyzed using multidimensional scaling and hierarchical cluster analysis. The Free Sorting Experiment was also analyzed using dual scaling. The results from all three methods were compiled and a set of 34 categories that proved to be stable across multiple methods of analysis was formed. A multidimensional perceptual image semantic space has been suggested and advantages to utilizing such a structure have been outlined. The 34 fundamental categories were represented by 10 perceptual dimensions that described the underlying perceptions leading to categorical assignments. The 10 perceptual dimensions were *humanness, artificialness, perceived proximity, candidness, wetness, architecture, terrain, activeness, lightness, and relative age.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

semantics, si-man'tiks, The study of meanings. The term is derived through Greek "semainein" ("to signify" or "to mean"). It is concerned with the relation between words or other symbols and the objects or concepts to which they refer, as well as with the history of meanings and the changes they undergo. (Semantics, 2001)

1. INTRODUCTION

The term *semantics* refers to the aspects of meaning that are expressed in a language. Photographic images tend to be perceived as belonging to broad categories of images such as *human portraits, landscapes, sports, animals, etc.* These categorical identifiers are referred to as *image semantics*—language that is used to describe the meaning of pictorial content.

Many tasks in imaging science are image-dependent. While a particular dependency might simply be a function of certain physical attributes of an image, often it is closely related to the perceived semantic category. Therefore, a thorough understanding of image semantics would be of substantial practical value. For example, there are many imaging tasks where image dependencies can influence the results of a certain processing task. Gamut mapping, halftoning, contrast adjustment, and compression are all dependant on the type of image being processed. Additionally, image quality judgments have been shown to exhibit image dependencies (Montag and Kasahara, 2001). If one could determine in advance what type of image was being processed, then an appropriate set of processing parameters could be selected so that the best possible result could be obtained.

Before one can identify which fundamental image category a particular image belongs to, it is necessary to determine what those categories are. There has been a lot work in the area of automatic image classification (Wardhani, 2003; Lee, et al., 2005; Le Borgne, et al., 2003). However, most of these methods rely on finding image descriptors that a machine can discriminate. Unfortunately these same descriptors are not necessarily perceptually significant. The difficulty lies in the fact that low-level image descriptors such as color and contrast fail to capture important semantic information, lack fine discrimination, and do not tend to match human perception (Mojsilovic, Hu, & Soljanin, 2002). Another concern is related to the basic goal of automatic image classification. "Classification pertains to a *known* number of groups, and the operational objective is to assign new observations to one of these groups." (Johnson and Wichern, 2002, p. 668). The problem here is that the fundamental groups, or categories, are *not* known and are simply selected by researchers based on a wide range of inconsistent criteria.

There is a body of work that addresses the problem of correlating image semantics with machine discernable image descriptors (Depalov, et al, 2006; Iqbal and Aggarwal, 2002; Le Borgne, et al., 2003; Lee, et al., 2005; Mojsilovic and Gomes, 2002; Mojsilovic and Hu, 2000; Mojsilovic, Hu, & Soljanin, 2002; Mojsilovic and Rogowitz, 2001a; Serrano, et al., 2002) but work in this area is not extensive. Further, just as there are image dependencies, it has been shown that any set of fundamental semantic categories that are determined will be influenced by the image genre (Laine-Hernandez and Westman, 2006). In other words, the categories that are identified will be directly dependent on what images were used in the study and images from different areas of photography will exhibit different fundamental categories.

Besides using semantic categories to optimize image processing algorithms, there are other ways this information can be utilized. Image retrieval from large databases of images is a significant problem and techniques that can help the user identify the image they are looking for more quickly can greatly improve the process. Much of the image classification research has been oriented toward solving this problem (Chen, et al., 2005) (Cox, et al., 2000; Chen, et al., 2003). Unfortunately, much of this work does not directly address the problem of first identifying the semantic categories that the images should be divided into. Instead, categories are selected based on either intuition or the ease of relating the image to low-level descriptors. Little or no regard for the perceptual significance of the categories is considered.

Another use for semantic image categories is image quality research in which the results can be shown to be image dependent. Because the image quality resulting from an image processing algorithm or an imaging system will depend on the image being evaluated, it would be very useful to know in advance what the fundamental image categories are. Although this issue is not unknown to researchers, most often the solution is to pick a variety of images arbitrarily to include in an experiment or study. There are two potential problems with this approach. First, it is possible that there are whole image classes that have been inadvertently left out of the study thereby making the results less comprehensive than desired. The second problem involves including multiple images that represent the same category which can result in unnecessarily duplicating effort. This could result in extra work (particularly in collecting the data) which would not produce any additional information. Therefore, by knowing in advance the fundamental image categories for a particular application, one can be sure that they are testing the full range of possible image types without unnecessary duplication of effort.

By identifying the fundamental image categories for a particular application, one can apply this knowledge to develop optimized image processing algorithms, better image retrieval systems, and design more targeted image research studies.

2. EXPERIMENTAL

The primary goal of this research was to identify the fundamental image categories for typical consumer imagery (snapshots). To achieve this goal, two psychophysical experiments were conducted. Experiment I was a Free Sorting Experiment and Experiment II was a Distributed Experiment which was conducted over the internet.

2.1 IMAGE SELECTION PROCESS

The design of both psychophysical experiments began with image selection. Because the same set of images was used for both experiments and because the usefulness of the experimental results is dependent on the images selected for inclusion in the studies, a great deal of care was taken to ensure that the images were representative of the intended application and that they spanned a wide gamut across potential image categories. There is an inherent difficulty in this process because it is necessary to first make categorical judgments in order to select a range of images for the studies, but determining categorical judgments is precisely the objective of the research. Therefore it was necessary to approach image selection with the most objective methods possible.

2.1.1 Category Selection

Without knowing the image categories in advance, how does one determine that the widest range of categories are properly represented and how does one avoid introducing bias by pre-filtering according to a personal observation? The solution was to make use of results from prior studies as a starting point. An extensive list of semantic categories was compiled from many sources (Mojsilovic and Rogowitz, 2001b; Wardhani, 2003; Wardhani and Thomson, 2004; Oldfield, 2005; Rogowitz, et al., 1998) and is listed in Table 2-1. There were many studies investigated that are not included in Table 2-1, however the categories that were identified in those studies were already represented and therefore, there was no need to report redundant information. Categories that had been previously identified by other researchers but were not considered semantic categories—such as *shape dominant* and *geometric objects*—were not included in the final category selection. This is justified because the categories that were excluded were originally identified to represent objects described by various low-level image descriptors that were calculated from the test images rather than to represent image semantics. Also excluded were any categories that could not be considered typical consumer imagery such as product photography and artificial scenes (i.e. graphics). The final list of categories that were used in the image selection process are listed in Table 2-2.

TABLE 2-1: Categories from previous studies.

STUDY 1	STUDY 2	STUDY 3	STUDY 4	STUDY 5	STUDY 6	STUDY 7
Portraits	Landscape	Natural Scene	Manmade	People	High-Frequency	Animals
People Outdoors	People	People	Natural	Non-People	Low-Frequency	People
People Indoors	Shape Dominant	Geometric Shapes	More Human-like	General Occasion	Color	Indoor Scenes
Outdoor Scenes with People	Color Dominant	Single Object	Less Human-like	Vacation	High Key	Nature
Crowds of People	Texture Dominant	Multiple Objects		Special Days (Birthdays, Celebrations)	Low Key	Buildings
Cityscapes	Structure Dominant	Mainly Smooth		Weddings		Textures
Outdoor Architecture		Mainly Textural		Holidays		Manmade
Technoscenes				Recreation, Sporting Events		
Objects Indoors				Subject Distance <10 ft.		
Waterscapes with Human Influence				Subject Distance 10-30 ft.		
Waterscapes				Subject Distance >30 ft.		
Landscapes with Mountains				Outdoors		
Sky/Clouds				Indoors		
Winter/Snow				Electronic Flash		
Green Landscapes				Sunlight/Daylight		
Landscapes with Fields and Foliage				Cloudy Day		
Plants, Flowers, Fruits, and Vegetables				Dawn/Dusk		
Animals				Indoor Natural Lighting		
Textures, Patterns, and Close-ups				Indoor Tungsten		
				Indoor Fluorescent		
				Indoor Mixed Lighting		
				Backlit		
				Low Light		

TABLE 2-2: Final categories used as a criteria in the image selection process.

Categories for Image Selection Process
Portraits
People Outdoors
People Indoors
Outdoor Scenes with People
Crowds of People
Cityscapes
Outdoor Architecture
Objects Indoors
Objects Outdoors
Waterscapes
Waterscapes with Human Influence
Landscapes with Mountains
Landscapes with Fields and Foliage
Green Landscapes
Landscapes with Water
Sky/Clouds
Winter/Snow
Plants, Flowers, Fruits, and Vegetables
Animals
Textures, Patterns, and Close-ups
General Occasion
Vacation
Special Days (Birthdays, Celebrations, etc.)
Weddings
Holidays
Recreation
Sporting Events
Nature

2.1.2 Image Selection

There were four criteria used in the image selection process. These were based on criteria identified from prior work and were determined to be a reasonable guide for image selection. They four criteria used were:

1) *Wide range of categories* — For each category listed in Table 2-2, images were selected such that each category was represented by a minimum of four images. Naturally, there was an inherent overlap and some images could be considered to fill the requirements for more than one category. For example, if one category is *people outdoors* and another is *nature scenes*, then an image of a person in nature could fulfill the requirement for both categories. Many categories were represented by far greater than four images.
2) *Camera zoom* — For each category, a range of wide-angle, normal, and close-up images were included.
3) *Image orientation* — A distribution of landscape and portrait orientation was included for each of the categories.
4) *Color* — A broad range of color and lightness levels was included. These were determined by examining average CIELAB values for each image, which was calculated as a simple average of every pixel color. Care was taken to include an even distribution in all three dimensions — L*, a*, and b*.

The images were obtained from a variety of sources. Most of the images in the final selection were supplied by Lexmark International, Inc. A library of nearly 1,000 images were initially provided. Despite this large number of images, there was not enough variety among them to satisfy all four selection criteria. After pre-editing the images to determine what additional images were required to satisfy all of the criteria, the remaining images were obtained from the Corel Image Database and from the Munsell Color Science Laboratory. For example, after an analysis of average CIELAB values for the selected images, it was determined that there were no *very light* images which was an indication that pictures with snow were missing. Therefore, a variety of *snow* images were added from the Corel Image Database.

A total of 321 images were included in the final selection. The complete set of 321 images can be found in Appendix A. To the experienced scientist, this may seem alarming since the typical number of images included in this type of study is usually closer to 100 or fewer. The reason for using a smaller number of images is because, for many experimental designs, as the number of samples (images) increases, the number of required observations increases rapidly. This generally makes a sample size of 321 infeasible to work with. The experimental design for the Distributed Experiment (which would be more sensitive to sample size than the Free Sorting Experiment) was taken into consideration while deciding on the number of images to include. The details of how the experimental design would effect selection of a sample size is described in section 2.4.1.

2.2 IMAGE PREPARATION

The 321 test images were obtained from a variety of sources. Some of the images were tagged with an ICC profile, but many were not. For this reason, the images needed to be adjusted in order to present a consistent appearance to the observers. Experiment I required printed output of the images and Experiment II required the images to be viewed over the internet. A typical strategy for preparing images for the internet is to use the sRGB color space. Therefore, all test images were adjusted so that they produced a pleasing image while viewed in the sRGB color space.

The first task was to characterize the monitor on which they would be evaluated. This was accomplished using industry standard calibration hardware and software. An Apple LCD display was calibrated using a GretagMacbeth EyeOne Pro spectrophotometer and GretagMacbeth ProfileMaker Pro software version 5.0. Images were viewed in Adobe Photoshop. A Lexmark inkjet printer was used to make prints. The printer was characterized using ProfileMaker Pro and a GretagMacbeth Spectrolino Spectroscan spectrophotometer.

If an image was untagged, it was assigned the sRGB profile. If the image was tagged with something other than sRGB, then it was converted to sRGB. All images were then adjusted, when necessary, so that they produced a pleasing appearance while viewed under the conditions specified by sRGB. Since the objective of this study was not dependent on specific colors in the images, it was not necessary to ensure exact color reproduction. The purpose for setting up color management and performing color adjustments was simply to normalize the appearance of images from a variety of sources.

Once all images were adjusted, they were scaled to a common size appropriate for the internet (200 pixels on the short dimension and 300 pixels on the long dimension) and saved in the JPEG format using low compression (Photoshop level 10) which yielded file sizes ranging from 52KB to 120 KB. Some images were a different dimension and were cropped to match the 2 x 3 aspect ratio. Next they were printed on the calibrated Lexmark printer at a size of 4 x 6 inches. Finally, the images were trimmed and numbered with bar codes on the back to aid the data collection process.

2.3 EXPERIMENT I — FREE SORTING EXPERIMENT

The first experiment was a tabletop sorting experiment. Observers were asked to sort the 321 test images into piles that represented categories. They were free to create as few or as many piles as they felt necessary to properly represent the categories. They were also free to change their mind and rearrange images in the piles as needed. Exact instructions as to how to categorize the images were not provided. The instructions provided to the observers read as follows:

> You will be given a stack of 321 4x6 photographic prints. Your task is to sort these into piles that represent different categories of image types. You may decide by what criteria to separate the images into categories and you may create as many piles (categories) as you feel are necessary.
>
> If a particular image seems appropriate for more than one category that you have defined, then use whichever identifying feature you feel is the primary feature, or create a new category. If you create a new category, remember to go through the images that have already been sorted to see if any of those belong to the new category.
>
> After you are finished, you will be asked to complete one additional small task.
>
> Thank you!!!

After the observers completed the task of separating the images into piles, they were asked to write down what criteria they used to separate the images into categories. Finally, the observers were asked to name the categories that represented each pile that they made.

2.3.1 Observer Statistics

Thirty (30) observers participated in the sorting experiment of which there were 19 males and 11 females with an average age of 38 years old. Twenty-four observers were considered expert observers. Seventeen observers were from the United States of America, eight observers were from China, two observers were from Japan, two observers were from Iran, and one observer was from India.

2.4 EXPERIMENT II — DISTRIBUTED EXPERIMENT

The second experiment was a distributed experiment conducted over the internet. Because of the nature of the experiment, it was necessary to first obtain Institutional Review Board approval from Rochester Institute of Technology's Human Subjects Research Office.

2.4.1 Experimental Design

The method of triads was used to collect similarity and dissimilarity data. Because the method of triads requires $\frac{n(n-1)(n-2)}{6}$ observations per observer where n is the number of images in the experiment, for 321 images and 30 observers this would require 163,838,400 individual observations. This is clearly unreasonable and is also the main reason most other studies limit the number of images to an average of one hundred or fewer. Therefore, alternate methods were explored that could help reduce the number of total required observations.

By limiting the number of samples that are to be compared with one another, the *incomplete block design* can greatly reduce the work that needs to done in an experiment. However, for 321 images there would be approximately 1.3 million total observations required and even this would still result in an intractable experimental design.

The method selected for this experiment was *Non-Repeating Random Paths* (Moroney and Tastl, 2005) which only requires n observations per observer where n is the number of images in the experiment. However, because this method generates a sparse matrix, the number of observers must be increased by an order of magnitude. Therefore, the total number of observations required becomes $300n$ which is equal to 96,300 observations for n=321 images. With a distributed experiment intended to reach a large number of people over the internet, it is assumed that not many people would choose to participate if the experiment was not quick to complete. One cannot expect to find observers willing to make 321 judgments. Therefore, the total number of observations was divided such that each observer would only be responsible for 10 judgments. In other words, the experiment must reach 9,630 people who will each judge 10 sets of triads per session. This is now an attainable goal.

One might question the validity of dividing each set of 321 observations intended for a single observer into smaller sets intended for 32 observers. The decision to do this is justified because spreading the observations over a greater number of observers will have the effect of removing any bias that might have been introduced by a single observer. Moroney (2003) describes a "*distributed experiment* in which the time requirements for each observer is reduced to a minimum by having a large number of observers, none of which complete the entire experiment. This reduces the impact of any given participant and provides a means to reduce the effect of multiple submissions and disruptive observers." Because observers in the present study were encouraged to repeat the experiment as many times as they desired, it is possible that some amount of observer bias is re-introduced, but this is considered a reasonable risk when evaluated against the total number of observations.

A comparison of the total number observations required for each of the three experimental designs is given in Table 2-3. This table was useful in helping to decide how many images to include in the experiments, as discussed in section 2.1.2.

TABLE 2-3: Total number of required observations for n samples.

n	100	200	300	400	500	600
Method of Triads	4,851,000	39,402,000	133,653,000	317,604,000	621,255,000	1,074,606,000
Incomplete Block Design	151,500	603,000	1,354,500	2,406,000	3,757,500	5,409,000
Non-Repeating Random Paths	30,000	60,000	90,000	120,000	150,000	180,000

Non-Repeating Random Paths (Moroney and Tastl, 2005) is a technique that reduces the burden for each observer by only requiring a random path to be evaluated. A random path is defined by taking each of the *n* samples and randomizing them. For the method of triads, the first set of three consecutive samples is evaluated. Utilizing a moving window that is three samples wide, the window is then stepped down the path by one sample and the next three samples within the window are evaluated. This continues 321 times until all sets of three consecutive samples have been viewed. Samples at the ends of the path are wrapped around to complete the path.

2.4.2 Experiment Interface

Observers were first presented with a welcome screen and an instruction screen that described how to perform the experiment. To entice people to take the experiment, a drawing for a free Apple iPod was held where each time the experiment was completed successfully, the participant would receive one entry into the drawing. Observers were encouraged to take the experiment multiple times. Figures 2-1 and 2-2 show the welcome screen and instruction screen respectively.

Observers were presented with three images at a time and were asked to pick which two images were the most similar and which two images were the most dissimilar. A sample experiment page demonstrating the interface is shown within Figure 2-2. After making the selections, the next set of three images was presented and this continued until all 10 triads had been viewed. At the end of the experiment, the observers had the option to enter comments and to repeat the experiment with a different set of images.

The interface was first prototyped in Matlab to validate the experimental design. The final implementation was created using basic HTML pages that were generated and customized by use of PHP. The data was collected into a MySQL database. In addition to the similarity and dissimilarity judgments, additional information about each session was recorded and is reported in Table 2-4.

FIGURE 2-1: Distributed Experiment welcome screen.

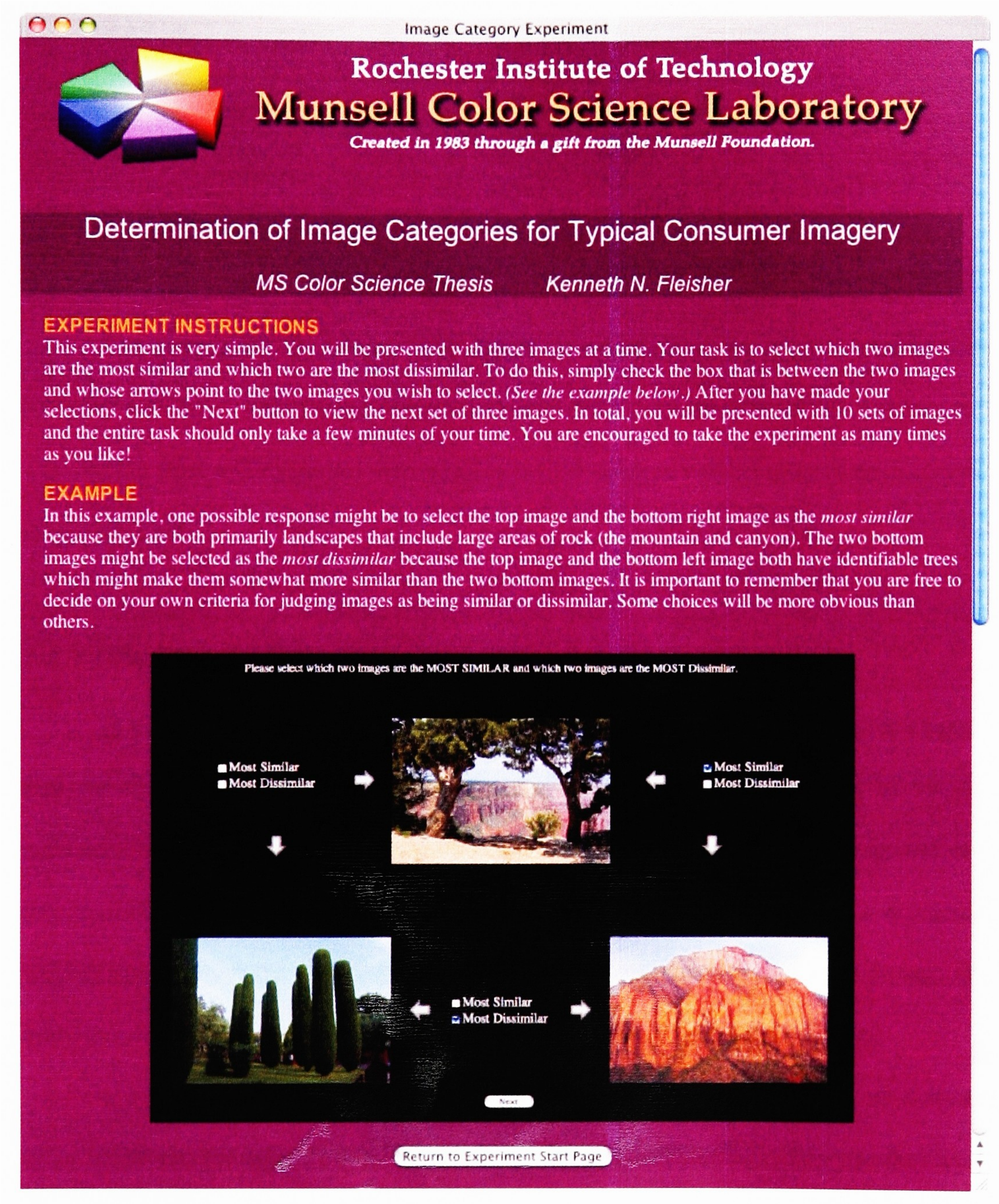


FIGURE 2-2: Distributed Experiment instruction screen.

TABLE 2-4: Metadata collected during Distributed Experiment.

Data Collected Automatically	Data Collected Voluntarily
Session ID	Gender
Session Start Time	Age
Session End Time	Comments
Session IP Address	
Session Agent ID (Browser)	

The session ID, IP Address, and Agent ID information was used to troubleshoot potential problems with internet connections and with the execution of the experiment. The session start and end times were used to help determine valid responses as discussed in section 2.4.4.

2.4.3 Observer Statistics

A total of approximately 9,152 people participated in the experiment which began data collection on March 13, 2006 and ended on May 17, 2006. A total of 98,364 individual trials were collected. An individual trial included one similarity judgment and one dissimilarity judgment. Because participation over the internet introduces a decreased level of control over the experiment versus an experiment conducted in a lab setting, the trials were pre-filtered for valid results to reduce the amount of noise in the data.

If a participant entered any of the voluntary information, then they were counted as an observer. Participants who included their e-mail address were only counted once even if they revisited the experiment at a later time. The goal was to count how many different people participated in the experiment. However, participants who did not enter their e-mail address remained anonymous and it was not possible to determine if they returned to the experiment at a later time. Therefore, the actual number of observers may be somewhat lower than what is reported. Additionally, participants who did supply any of the voluntary information were not counted as observers with respect to observer statistics. Table 2-5 list the observer statistics that were collected for the Distributed Experiment. There were 5,195 participants who provided voluntary information.

TABLE 2-5: Observer statistics for the Distributed Experiment.

	Gender		Age (years)
# Responded	5,108	# Responded	4,905
Male	3,257	Average	30.8
Female	1,851	Median	24
		Maximum	85
		Minimum	8

2.4.4 Pre-Filtering Observer Responses

To pre-filter the 98,364 trials and identify invalid responses, it was necessary to make certain assumptions. The first assumption was that most observers would be participating in order to try and win the drawing for a free iPod. Therefore, the primary reason to provide invalid responses would be to complete as many sessions as possible (a session consisted of 10 individual trials) in order to increase the odds of winning the drawing. For this reason, only users who completed 10 or more sessions were evaluated for invalid responses. If there were users with fewer than 10 completed sessions who provided invalid responses, then the impact on the results should be small. In order to be entered into the drawing, the observer was required to include their e-mail address. It is possible that someone did not include their e-mail address and still had greater than 10 sessions, but there was no mechanism to identify such a case. It is also unlikely that someone would go to the trouble of disrupting the experiment without the possibility of winning the drawing, so this case was not considered. Of the 4,155 participants who entered their e-mail address, only 84 participants completed 10 or more sessions. The sessions corresponding to the 84 participants were evaluated based on five possible situations:

1) *Scripted responses* — If a script were created to complete many sessions without the need to actually participate, then the sessions would have been completed very quickly. On average, it took between 2 and 3 minutes for a participant to complete the experiment honestly but a script would likely complete each session very quickly. Therefore, a conservative cutoff was to remove any sessions that were completed in less than 60 seconds. There were 3,222 trials from 323 sessions completed in less than 60 seconds that were removed. *(The number of trials was not equal to 10 times the number of sessions due to some PHP code errors which caused some incomplete sessions to be recorded. The problem was identified and fixed within the first few days of the experiment. Individual trials that were recorded during this time are valid even though the sessions were incomplete.)*
2) *Same response* — A fast way to enter responses and increase the number of completed sessions would be to blindly select the same response for every trial. It is statistically unlikely that these would be valid responses and therefore a series of sessions from a single user with all *same responses* would be a candidate for exclusion.
3) *Patterned response* — An approach similar to the *same response* approach would be to enter a patterned response to all trials, such as A, B, C, A, B, C, ... Patterned responses were excluded from the analysis.

4) *Random response* — This was the most difficult to identify because it required the author's judgment as to whether the response was random or sincere. For all 84 users with greater than 10 sessions, a random viewing of their responses was conducted. Sometimes it was undeniable that the response was not sincere. For example, if image 1 was a dog, image 2 was also a dog, and image 3 was a house and the observer selected images 1 and 3 as the most similar, then this can be considered an invalid response. However, a very conservative approach was taken so as not to inadvertently exclude valid responses where the judgment was simply different from the author's. Only after many invalid trials were identified for a particular observer were that observer's responses excluded.

Based on *same response*, *patterned response*, and *random response* criteria, 510 trials were excluded from the analysis.

5) *Negative comments* — There was an option at the end of the experiment to include comments. While most comments were either positive or inquisitive, there were four sessions where the participant simply stated that they did not even look at the images and that they simply responded randomly. Naturally, these sessions were excluded from the analysis.

To summarize, there were 98,364 individual trials collected. 3,222 trials were excluded due to a total session time of less than 60 seconds. 510 trials were excluded due to *same*, *patterned*, or *random responses*. 40 trials were excluded based on user comments. The remaining 94,592 individual trails were included in the data analysis.

3. METHODS OF ANALYSIS

To analyze the experimental data, several methods of analysis were used. Although a complete description of each method is beyond the scope of this report, each method is briefly described and key features are explained. For a full treatment of these techniques, please refer to the relevant references.

3.1 MULTIDIMENSIONAL SCALING

"The problem of multidimensional scaling, broadly stated, is to find *n* points whose interpoint distances match in some sense the experimental dissimilarities of *n* objects" (Kruskal, 1964a). Another way to express this is to say that multidimensional scaling "enables us to represent the similarities of objects spatially as in a map." (Schiffman, et al., 1981, p. 3) Thus the primary output of multidimensional scaling is a low-dimensional, spatial representation of points where each point corresponds to an object in the original data. This configuration, or ordination of the data (Johnson and Wichern, 2002, Ch. 12), is then interpreted in an effort to uncover the organizing concepts and underlying dimensions that are being investigated. This last statement is of particular importance because it implies that dimensionality and significant characteristics of the objects need not be known *a priori* and are discovered as a result of interpretation of the configuration.

As input, multidimensional scaling requires only similarity (or dissimilarity) data. When the data is defined by an interval or ratio level of measurement, the algorithm is called *metric multidimensional scaling*. When the data is defined by only an ordinal level of measurement, the method is called *nonmetric multidimensional scaling* (Young and Hamer, 1987, Ch. 2).

Multidimensional scaling is a family of algorithms that operate on the principle of minimizing the error between similarities (or dissimilarities) in the experimental measurements and distances in the configuration. There are different measures available to minimize in the objective function. One of the most common measures is a quantity called *stress* and is given in Equation (1). Stress is "essentially the root-mean-square residual departure" from the hypothesis that "the observed dissimilarities differ from the true dissimilarities only because of random fluctuation." (Kruskal, 1964b)

$$stress = S = \sqrt{\frac{\sum_{i<j}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}} \tag{1}$$

where d are the dissimilarities and $\hat{d}$ are the "numbers which minimize S subject to the constraint" of monotonicity (Kruskal, 1964a). A variation of stress known as *sstress*, or squared stress, is given in Equation (2). Sstress is another common measure of how well the configuration fits the data (while some authors use the sum of the disparities to the 4th power for normalization, the implementation of Matlab used in this analysis normalizes with the sum of the distances to the 4th power).

$$sstress = \sqrt{\frac{\sum_{i<j}\left(d_{ij}^2 - \hat{d}_{ij}^2\right)^2}{\sum_{i<j} d_{ij}^4}} \tag{2}$$

3.1.1 Goodness-of-Fit

Stress and sstress are essentially measures of the goodness-of-fit for the configuration. A perfect fit will have a stress of zero, but this is unlikely to ever happen with experimental data. As the fit worsens, the stress score increases. Therefore, it is more proper to think of this as a measure of badness-of-fit. However, because of the general acceptance of the term *goodness-of-fit*, we will continue to use this expression. Table 3-1 enumerates a general "rule of thumb" for interpreting the stress value as it relates to the goodness-of-fit (Kruskal, 1964a).

TABLE 3-1: Goodness-of-Fit "rule of thumb" for stress values.

Stress	**Goodness-of-Fit**
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent
0%	*Perfect*

Another technique for investigating the goodness-of-fit is to examine a scatter diagram, also known as a Sheppard diagram. A scatter diagram is "a plot comparing the distances derived by [multidimensional scaling] and the transformed data (disparities) with the original data values or proximities." (Schiffman, et al., 1981, p. 17). What the scatter diagram can tell us is whether the stress value is reliable, if there is degeneracy, and whether the method of computing the configuration is appropriate.

One thing to look for in a scatter diagram of a metric solution is how clearly the points fit some curve. When the points seem to fit some function other than the function *f* assumed in the model (such as a linear function), then the stress value will be magnified due to an incorrect assumption about the function *f*. This may be an indication that the data needs to be reanalyzed. For a non-metric solution, one can only check to see if the function is monotonic. If it is non-monotonic then it may be necessary to try a different starting position. The general shape of a smooth curve drawn through the points can sometimes provide information about the data, such as artifacts due to a particular data collection method. Lastly, if the data points on a scatter diagram are strongly clumped, this is a good indication of degeneracy. "If degeneracy occurs, the clustering it springs from should be noted and considered, but no other conclusions should be drawn. In particular, the very small stress should not be taken as indicating good fit in a substantive sense, since it is obtained by violating two tacit assumptions: that the true relationship between distance and proximity is smooth, and that points should only lie in the same position if the corresponding objects function as virtually identical." (Kruskal and Wish, 1978, pp. 29-30).

3.1.2 Dimensionality

It is possible to estimate the underlying dimensionality of the data. One way to approach this is to compute the stress (or sstress) value for each of several configurations representing a range of dimensions. By examining a plot of dimension versus stress, it may be possible to find some clues about dimensionality. As the number of dimensions increases, the stress value should decrease. In many cases, there will be an elbow in the curve at the dimension that represents the true dimensionality. Figure 3-1 illustrates this concept where the true dimensionality of the data is three dimensions. However, not all data will produce such a clear indication of dimensionality. When that happens, it may be possible to interpret a range of possible dimensions or it may not be possible to interpret dimensionality at all.

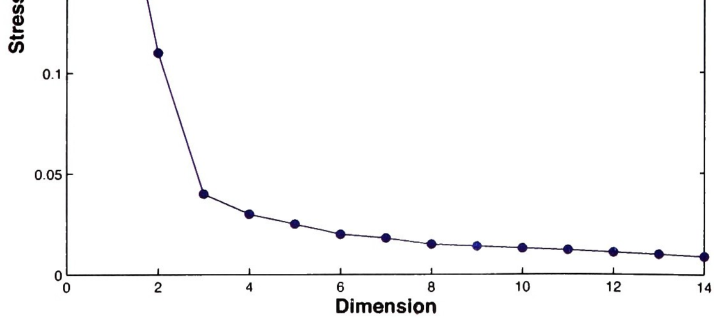


FIGURE 3-1: Example of Dimension vs. Stress plot suggesting 3-dimensional data.

3.1.3 Configuration

The primary output of multidimensional scaling is a low-dimensional configuration of points that graphically represents similarities in the data. By examining the configuration, it is possible to look for clues that help to interpret the organizing concepts inherent in the data. Although the configuration will attempt to produce the best spatial representation of the data, the orientation of the configuration is arbitrary. Take for example Figure 3-2 which demonstrates the two-dimensional configuration resulting from multidimensional scaling of the distances between ten cities. It is clear that the dimensions north/south and east/west are not in alignment with the axes. The configuration is still correct because the distances between the points would remain the same if the plot were rotated so that the dimensions did align with the axes (Kruskal and Wish, 1978, Ch. 2). It is also possible that the dimensions are not orthogonal to one another and that there may be only one dimension, or potentially more than two dimensions that are interpretable on a particular 2-dimensional configuration.

A configuration can exhibit local structure as well as global structure. By selecting a cluster of points within the global configuration and performing multidimensional scaling on just these objects, the new configuration may reveal additional structure that was previously obscured. This type of neighborhood interpretation can lead to further understanding of the organizing concepts and underlying dimensions.

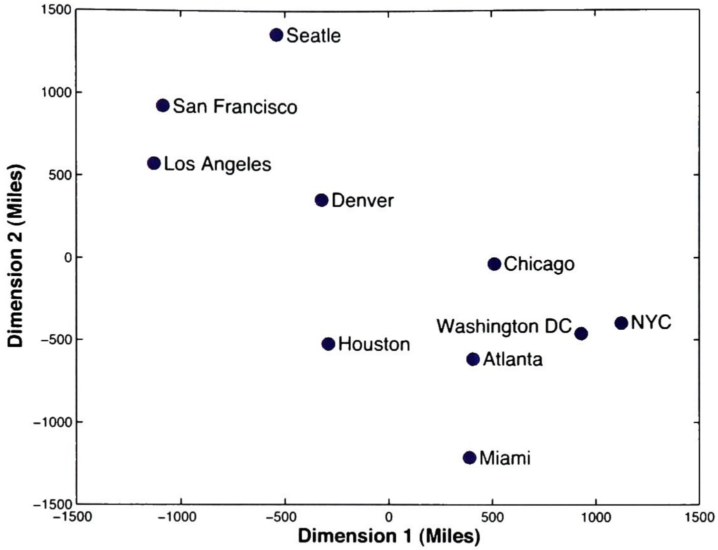


FIGURE 3-2: Example of a 2-dimensional multidimensional scaling configuration of distances between 10 cities. The orientation is arbitrary.

3.2 DUAL SCALING

Dual scaling is often referred to as principal component analysis for categorical data (Maraun, et al., 2005). Principal component analysis operates by maximizing the variance for a set of variables, or principle components, which are linear combinations of the original variables. The principal components are orthogonal to one another which minimizes redundancy of information. Dual scaling accomplishes the same goal except that it operates on categorical data to produce multidimensional decomposition of the data as explained by Nishisato & Nishisato (1994, p. 8):

> When the optimal solution does not explain the data in an exhaustive way, dual scaling determines a second set of scores and weights that maximally explains the portion of data unexplained by the first optimal solution. If the data cannot be perfectly reproduced by the first two sets of optimal

scores and weights, dual scaling looks for the third optimal set of scores and weights that maximally explains the portion of the data left unexplained by the first two solutions. This process continues until the original data can be perfectly reproduced by the solutions obtained so far, that is, until the data are exhaustively analyzed. This process is called *multidimensional decomposition* of data.

This type of decomposition identifies the solution that provides the most information first, followed by the solution that provides the next largest amount of information, and so on. This process makes it possible to represent the most information with the fewest number of solutions. The usefulness of the solutions can be determined by the amount of variability of the original data that is included in the solutions and by the researcher's ability to interpret the solutions.

There are many aspects of dual scaling that differentiate it from other types of analysis. Some of the key characteristics of dual scaling are listed by Nishisato (1994, Ch. 2) and are summarized here:

1) Dual scaling provides a simpler, often clearer, description of data, thus serving as a technique to form a useful summary of otherwise complex data.
2) It derives a numeric (quantitative) description from non-numeric (qualitative) data...
3) It handles analysis of a variety of so-called categorical data...
4) It offers an exhaustive analysis of information in the data, often through *multidimensional analysis*...
5) It serves as a technique for *discriminant analysis* of categorical data.
6) It extracts information from data in optimal ways (e.g., derives test scores which have maximal reliability).

7) It uses individual differences in judgment to explore the data, rather than averaging them out as in most statistical analyses. Individual differences are often more interesting than average responses.
8) It can quantify qualitative information so that traditional analysis (e.g., analysis of variance) for quantitative data may be carried out.

There are numerous techniques for analyzing the results of dual scaling. One very useful outcome is the ability to visualize high-dimensional data. Humans normally have difficulty visualizing greater than three dimensions. Multidimensional scaling provides a way to represent high-dimensional data in a low-dimensional configuration as described in section 3.1.3. However, there is ultimately a loss of information when performing dimensional reduction. The dimensions that result from multidimensional scaling are not orthogonal and therefore may contain redundancy of information. This is where dual scaling can provide some insight where multidimensional scaling cannot. Because the solutions of dual scaling are orthogonal, each successive solution can be considered an added dimension with new information. Although it is still not possible to create a high-dimensional graphical representation of this information, it is still possible to evaluate dimensions greater than three.

By examining the objects at either end of a particular dimension, it is possible to develop an interpretation about certain characteristics of the objects which may lead to clues about the nature of the variance contained in that dimension. In other words, if the objects at opposing ends are significantly different in some way, identifying what that difference is will provide some information about the underlying structure in the data. For example, Figure 3-3 shows eight images at either extreme of the fourth dimension from a dual scaling analysis of image categorization. The eight images from the one extreme are all historic buildings. The eight images from the opposite extreme are comprised of land, water, and air vehicles. In other words, several modes of transportation vehicles are represented.

FIGURE 3-3: Example of evaluating the nature of variance in the fourth dimension of a dual scaling analysis. Even without a spatial configuration, it is possible to interpret characteristics of this dimension.

This example illustrates an important point about performing this type of analysis—the characteristics at the extremes of a particular dimension do not have be related to each other nor do they have to be opposites of one another (such as light and dark). Transportation vehicles are not the *opposite* of historic buildings. These two categories of images simply provided the greatest amount of variance for this particular dimension using this particular dataset. When the dimensionality becomes high enough that the amount of variance represented by that dimension is very small, then it will no longer be possible to derive a reasonable interpretation for the objects at the extremes.

"The chief aim of a dual scaling analysis is not statistical inference but rather the description of the high-dimensional categorical data structures that often arise in psychological research. The researcher who believes he or she has found an interesting relationship through the employment of dual scaling should, as per sound scientific practice in general, attempt to replicate the finding at a later date." (Maraun, et al., 2005) In the current study, this is accomplished through simultaneous analysis through multidimensional scaling and hierarchical cluster analysis.

A general description of dual scaling and some its uses have been presented. Although a thorough treatment of the mathematics of dual scaling is beyond the scope of this paper, an excellent source for the mathematical details can be found in Chapter 6 of Nishisato (1994).

3.3 HIERARCHICAL CLUSTER ANALYSIS

The basic objective of any cluster analysis algorithm is to discover natural groupings of the objects. Because it is not generally possible to explore all object groupings, it is necessary to use other methods that can find sensible clusters without the need to examine every cluster. Many clustering algorithms will group n objects into a fixed number of k clusters which must be specified in advance. Hierarchical cluster analysis does not have such a requirement and instead builds a tree-like structure, or hierarchy, which represents all values of k. This is achieved by using one of two general approaches—*agglomerative hierarchical methods* or *divisive hierarchical methods.* (Kaufman and Rousseeuw, 1990, Ch. 5).

3.3.1 Agglomerative Hierarchical Methods

Agglomerative hierarchical methods begin with computation of a similarity matrix between the objects. The initial clustering is defined as a single object per cluster so that there are the same number of clusters as objects. Groups containing objects with the greatest similarities are merged to form the first distance level of the hierarchy (the initial grouping of individual objects is considered the zero distance level). This process continues iteratively until all objects are merged into a single cluster (Johnson and Wichern, 2002, Ch. 12). Most implementations of hierarchical cluster analysis, including the Matlab implementation used in this study, are agglomerative hierarchical methods.

3.3.2 Divisive Hierarchical Methods

Divisive hierarchical methods operate in reverse of agglomerative methods. After computing a dissimilarity matrix between the objects, all objects are grouped into a single cluster. The initial cluster is then divided into two clusters such that the dissimilarity between the two groups is maximized. In other words, the objects in one group have the greatest distance from the objects in the second group or are the *most dissimilar*. The process of dividing groups into subgroups continues until each group contains only a single object (Johnson and Wichern, 2002, Ch. 12).

Most software implementations of hierarchical cluster analysis do not include divisive methods due to the high computational overhead required. While an agglomerative method contains $\frac{n(n-1)}{2}$ possible combinations to evaluate, a divisive method will have $2^{n-1}-1$ possible combinations. Note that the agglomerative methods will grow quadratically as n increases while the divisive methods will grow exponentially. The number of possible combinations to evaluate for a divisive method can quickly "exceed the current estimate of the number of atoms in the universe" (Kaufman and Rousseeuw, 1990, pp. 253-254). Therefore, few implementations of divisive methods exist.

3.3.3 Linkage Methods

For the agglomerative hierarchical methods, the main difference between them is the manner in which linkage distance is calculated. *Linkage* describes the process by which groups are merged. There are seven distance measures, also called *linkage methods*, implemented in Matlab—*single, complete, average, weighted, centroid, median, and Ward's*. Everitt (1974, Ch. 2) provides a good description of how each linkage method is calculated (except for weighted linkage which is described in the Matlab documentation (The Mathworks, Inc., 2004)) which is summarized here:

1) *SINGLE* – This method, also known as the *nearest neighbor* method, uses the smallest distance between objects. For groups with more than one object, distance is defined as the distance between their closest objects.

2) *COMPLETE* – This method, also known as the *furthest neighbor* method, is the same as single linkage for the initial grouping. After the initial step, for groups with more than one object, distance is defined as the distance between their most remote pair of objects.

3) *AVERAGE* – This method defines distance as the unweighted average distance between all pairs of objects in the two groups being compared.

4) *WEIGHTED* – This method defines distance as the weighted average distance between all pairs of objects in the two groups being compared.

5) *CENTROID* – This method defines distance as the distance between the centroid of all objects in one group and the centroid of all objects in a second group. A disadvantage of this method becomes evident when the size of two groups being merged are very different because the centroid of the new group will be very close to that of the larger group. Characteristics of the smaller group are therefore potentially lost. Another potential problem is that the resulting cluster-tree might not be monotonic. For the centroid linkage method to be meaningful, the similarity matrix must contain Euclidean distances.

6) *MEDIAN* – To overcome the potential disadvantage of the centroid linkage method, the median method defines distance as the Euclidean distance between the weighted centroids of the two groups being merged. The potential of the resulting cluster-tree not being monotonic

remains. For the median linkage method to be meaningful, the similarity matrix must contain Euclidean distances.

7) *WARD'S* – "Ward proposes that at any stage of analysis the loss of information which results from the grouping of individuals into clusters can be measured by the total sum of squared deviations of every point from the mean of the cluster to which it belongs. At each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined." (Everitt, 1974, p. 15) For the Ward's linkage method to be meaningful, the similarity matrix must contain Euclidean distances.

3.3.4 Interpretation of the Cluster Tree

Once the hierarchical cluster tree is calculated, there are numerous ways to interpret the results. Some of these analysis techniques involve quantitative measures. However, as is the case with multidimensional scaling and dual scaling, some of the most useful techniques involve qualitative interpretation.

3.3.4.1 Cophenetic Correlation Coefficient

The cophenetic correlation coefficient is a measure of correlation between the distance information, Z, generated by the linkage step and the distance information, Y, generated from the similarity matrix as the pairwise distances between observations in the original proximity data. In other words, it "measures the distortion of [the hierarchical cluster tree], indicating how readily the data fits into the structure suggested by the classification." (The Mathworks, Inc., 2004) A cophenetic correlation coefficient of 1 indicates a perfect representation of the data. As the coefficient decreases, the quality of the clustering decreases.

3.3.4.2 Dendrogram

A dendrogram, or tree diagram, is a graphical display of the binary cluster tree hierarchy. Clusters are represented as branches in the tree structure. Nodes at which the branches merge are positioned along a distance (or similarity) axis indicating the level of the fusion. The dendrogram can be used to identify natural cluster divisions in the data. This is accomplished by comparing the relative height of links in the tree with the heights of neighboring links that occur below it in the tree structure. For example, in Figure 3-4 the topmost links are a large distance from the links below indicating that there is inconsistency in the linkage at those levels. However, good consistency can be observed beginning at approximately the 0.6 level suggesting that there are three natural groupings in this data. In other words, inconsistent links can indicate a boundary between natural cluster divisions in the data. Interpretations such as this can lead to an understanding about the nature of the underlying structure. It should be noted that the "intermediate results—where the objects are sorted into a moderate number of clusters—are of chief interest" as long as the clustering makes sense (Johnson and Wichern, 2002, p. 683).

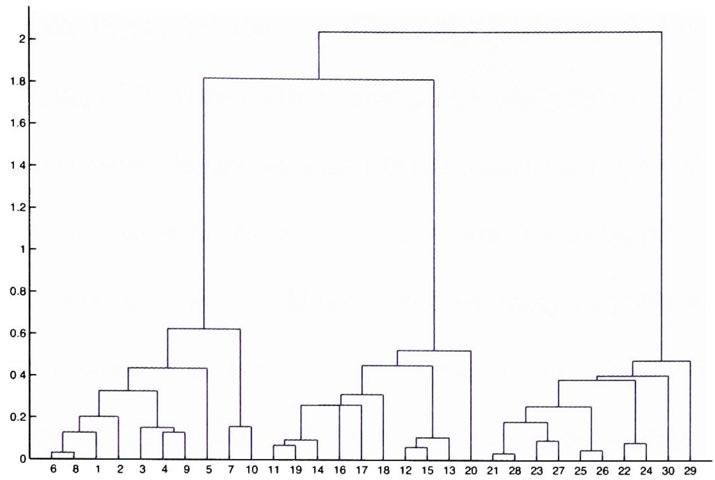


FIGURE 3-4: Example of a dendrogram. The large distance of the topmost links from the links below indicates inconsistency at those levels.

3.3.4.3 Silhouette Plot

A silhouette plot provides an alternate means for determining the natural grouping of objects and the underlying dimensionality. Although the silhouette plot is typically used with hard-clustering methods such as k-means and k-medoid which divide the objects into a fixed number of clusters, it is also possible to examine a silhouette plot for hierarchical methods by dividing the cluster tree into arbitrary clusters. The silhouette plot is a graphical display indicating how close each object of a cluster is with objects in neighboring clusters with values on the interval `(-1,1)`. Values close to 1 are very distant from neighboring clusters indicating a very good clustering. Values close to zero indicate objects that are not distinctly in one cluster versus another. Finally, values close to -1 are indicative of objects that have likely been classified to the wrong cluster.

For example, Figure 3-5 shows a silhouette plot of data that is known to have three natural groupings. The three clusters show excellent separation with most silhouette values near 0.9. In contrast, Figure 3-6 shows the silhouette plot where the same data has been divided into four clusters. Note how cluster two has a steeper slope with many silhouette values closer to zero. In addition, there are many negative values indicating that a four-cluster division is probably not the optimal division for this data. In practice, the distinction between a good and poor clustering may not be as easy to identify as in this simple example.

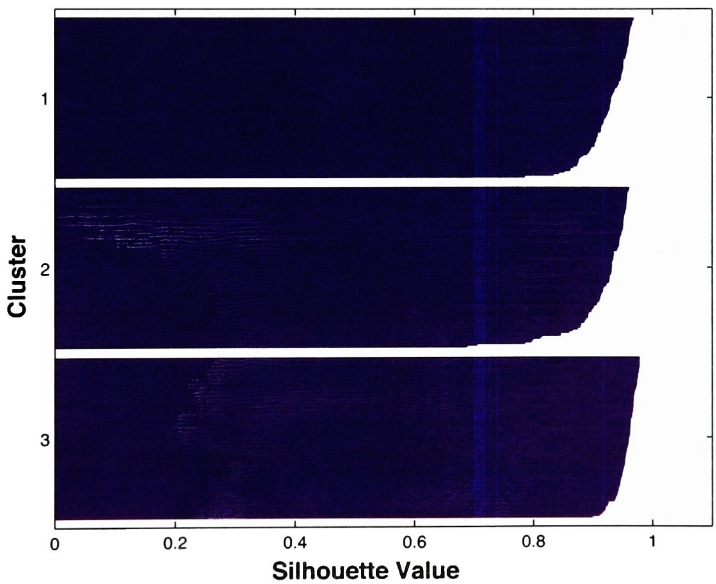


FIGURE 3-5: Example of a silhouette plot indicating three natural groupings.

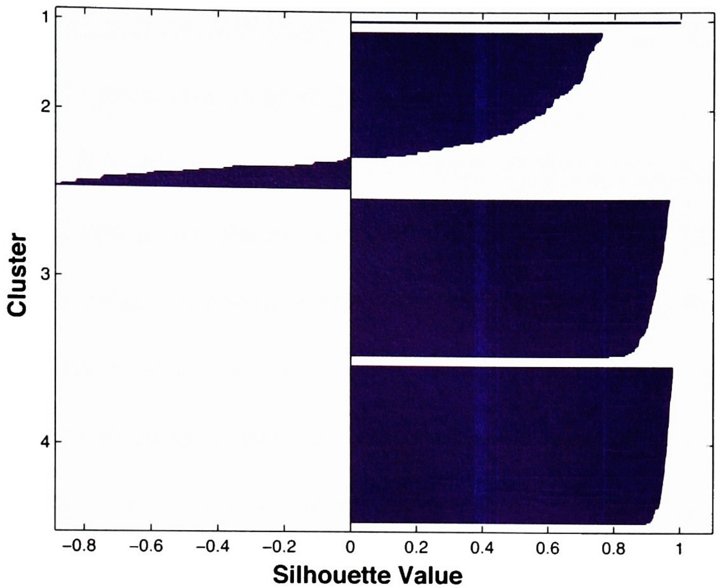


FIGURE 3-6: Silhouette plot showing 3-dimensional data in 4 groupings.

3.3.5 Cluster Interpretation

Visual examination of the individual objects that belong to each cluster identified by the methods described above can lead to an understanding of the structure of the data. If the basic objective of a cluster analysis algorithm is to discover natural groupings of the objects, then these groupings must be evaluated and interpreted. "For a particular problem, it is a good idea to try several clustering methods and, within a given method, a couple of different ways of assigning distances (similarities). If the outcomes from the several methods are (roughly) consistent with one another, perhaps a case for *natural* groupings can be advanced." (Johnson and Wichern, 2002, p. 693) For the present study, hierarchical cluster analysis was the only clustering method used, but the results were compared with the results of multidimensional scaling and dual scaling before making judgments about the natural groupings of the data.

Calculating the cluster tree using the seven distance measures described in section 3.3.3 and dividing the cluster tree into several sets of fixed clusters makes it possible to perform visual inspection of many possible configurations. For example, first compare the clusters from all seven distance measures for a fixed number of three clusters. If there are similarities between the resulting clusters from a majority of the methods, then those methods which exhibit the similarity can be considered stable for solutions for that fixed number of clusters. This comparison is repeated for four clusters, five clusters, etc., until the resulting groupings no longer make sense and cannot be interpreted. After the stable configurations are identified, then the individual objects contained in these clusters can be evaluated. Some questions to ask while examining the objects are "Do the groupings become more clear or less clear as the number of clusters increases?" and "Can any conclusions be drawn about the nature of the groupings?"

4. DATA PREPARATION

Before data analysis could begin, the raw data from the two experiments required some pre-processing. This was necessary because each method of analysis requires data to be in a particular format.

4.1 EXPERIMENT I — FREE SORTING EXPERIMENT

Data from the Free Sorting Experiment was formed into a dissimilarity matrix. Observers placed 321 images into piles according to similarity judgments. Each observer was free to make as many or as few piles as desired. A 321 x 321 zero-filled matrix M was formed. Each image was assigned a unique index ranging from 1 – 321 which was used as row and column indices into matrix M. For each observer, every time a pair of images was placed into the same pile, a value of 1 was added to matrix M at the coordinates corresponding to the indices of the two images. After all observations were tallied for all observers, the resulting entries were scaled by dividing by 30. Thirty observers represents the maximum number of times a pair of images could be sorted into the same pile. This scaled the results on the interval `(0,1)` and now represents a similarity matrix M_{sim}. To form the dissimilarity matrix M_{dis}, it was only necessary to subtract from 1 such that $M_{dis} = 1 - M_{sim}$.

4.2 EXPERIMENT II — DISTRIBUTED EXPERIMENT

After pre-filtering of the data as described in section 2.4.4, the observations were formed into a dissimilarity matrix. Experiment II collected both similarity judgments and dissimilarity judgments based on non-repeating random paths. A result of this method of data collection is that not every pair of images was evaluated the same number of times. Therefore, to form the dissimilarity matrix, first the results were tallied in the same manner as the Free Sorting Experiment. Every time a pair of images were judged as the most similar, the value at the coordinates corresponding to the indices of the two images was incremented by 1. The same procedure was followed for every pair of images that was judged as the most dissimilar to form a separate dissimilarity matrix. Because every pair was not observed the same number of times, it was necessary to keep a count c for each pair of images of how many times they were judged together. The similarity matrix was then subtracted from c, the maximum number of times that pair could have been judged as the most similar, on a pair-by-pair basis to obtain a dissimilarity matrix. The two dissimilarity matrices were then averaged to obtain a mean dissimilarity matrix. The mean dissimilarity matrix was then divided by the count c on a pair-by-pair basis to obtain the final frequency matrix (dissimilarity matrix).

5. DATA ANALYSIS

The Free Sorting Experiment and the Distributed Experiment both produced data representing similarity judgments. Once the data from each experiment was properly prepared, analysis proceeded using the same techniques for both experiments. The only exception was dual scaling which was only applied to the Free Sorting Experiment. This was because the nature of the data collected from the Distributed Experiment was not appropriate for dual scaling.

5.1 MULTIDIMENSIONAL SCALING

Analysis using multidimensional scaling was conducted. Because the data collected from both experiments was represented by an ordinal scale, non-metric multidimensional scaling was selected. The Matlab function *mdscale* was used to perform the calculations. Due to the large dataset, the time required for processing the higher dimensional configurations was unrealistic. Therefore, certain termination criteria were set. The function tolerance was set to *TolFun = 0.001* and the maximum number of iterations was set to *MaxIter = 600.* The criterion selected for the objective function was the sstress function because it tends to produce a smoother solution than the stress function (Kearsley, et al., 1995) and given the nature of the data collected from the Distributed Experiment (a certain amount of noise is expected), it was believed sstress would give a better result. The Free Sorting Experiment also used sstress to be consistent with the Distributed Experiment.

To verify that there were no adverse effects by selecting these criteria, several dimensions were calculated using a greater number of iterations and a smaller function tolerance. The results were compared from the two sets of criteria. Although the absolute value of sstress was very slightly shifted to lower values (Figure 5-1) , the trend was identical and nearly the same values were obtained indicating that the criteria that were used were valid. If the function criteria were harmful to the processing, then the stress values for the two conditions would be very different.

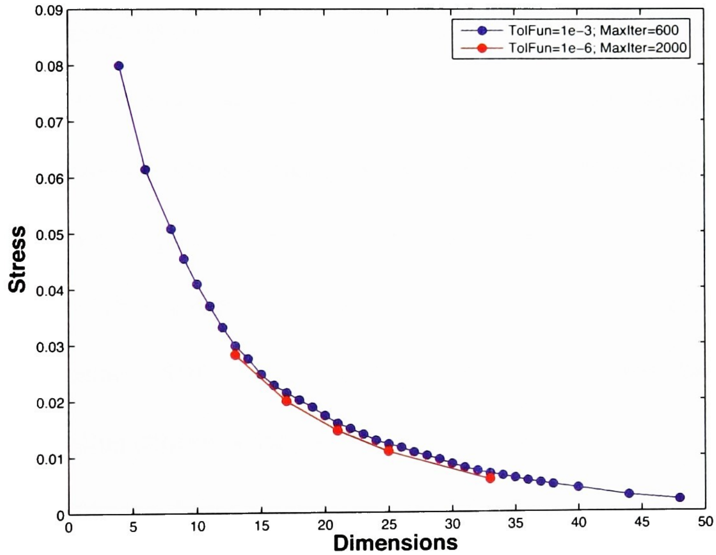


FIGURE 5-1: Validation of function criteria for multidimensional scaling.

5.2 DUAL SCALING

Analysis using dual scaling was conducted. Due to the nature of the data collection methods, only data from the Free Sorting Experiment was appropriate for this type of analysis and the Distributed Experiment was not analyzed with dual scaling.

Dual scaling accepts a similarity matrix as input. However, depending on the method of data collection the analysis must be conducted slightly differently. For the Free Sorting Experiment, the difficulty lies in the fact that each observer was permitted to make as many piles as determined necessary. The number of piles that one observer makes could be different from the number of piles that another observer makes. This is handled in dual scaling as a special case of multiple-choice data (Nishisato and Nishisato, 1994, Ch. 3). With multiple-choice data, there are a fixed number of options from which each observer selects a response. "In sorting data, the number of *piles* each *respondent* uses corresponds to the number of *options* of each *item* in multiple-choice data..." (Nishisato and Nishisato, 1994, p. 53). Except for this variation for the special case of sorting data, dual scaling proceeds normally.

One reason for the different treatment of sorting data is explained as follows: "... in sorting data, individual differences are revealed through subjects' unrestricted or free choices of piles, rather than the researcher's imposing decision on the number of piles. This distinction appears to explain the fact that dual scaling of sorting data often yields too many solutions to interpret. Therefore, the problem of how many solutions to adopt becomes more difficult with sorting data than with multiple-choice data or contingency tables." (Nishisato, 1994, p. 172).

Dual scaling was computed for the first 50 dimensions. For each dimension, the first eight and the last eight images were output for visual interpretation as described in section 3.2.

5.3 HIERARCHICAL CLUSTER ANALYSIS

Analysis using hierarchical cluster analysis was conducted. All functions used are part of Matlab's *Statistic's Toolbox* of which all are agglomerative hierarchical methods. Beginning with the original dissimilarity matrix of observations, the function `pdist` was used to obtain a dissimilarity matrix using the Euclidean distance metric. Each of the seven linkage methods described in section 3.3.3—single, complete, average, weighted, centroid, median, and Ward's—were then calculated using the function `linkage`. The cophenetic correlation coefficient was calculated for each of the seven resulting hierarchical cluster trees using the `cophenet` function.

Analysis of multidimensional data is often easiest to understand using graphical representations of the data. Each of the seven hierarchical cluster trees was formed into a dendrogram using the `dendrogram` function. When there are more than thirty data points in the original data set, a complete dendrogram can become very dense and difficult to interpret. When this occurs, as it does in the present study with 321 data points, the default behavior of the Matlab function `dendrogram` is to collapse some of the lower branches as necessary such that some leaves in the plot will correspond to more than one data point. This default behavior was applied to the dendrograms presented in this study.

Another graphical representation is the silhouette plot. As input to the function `silhouette`, it was necessary to first divide the data into a fixed number of clusters using the function `cluster`. It is useful to examine the silhouette plots for several linkage methods across several fixed cluster sizes. Therefore, the data was divided into a fixed number of clusters ranging from 2-15. Seven silhouette plots (one for each linkage method) for each of the resulting 14 groupings were then created.

Finally, clusters were calculated for 133 groupings (seven linkage methods for each of 2-20 fixed clusters). The resulting clusters were output as individual images for visual interpretation as described in section 3.3.5.

6. GOODNESS-OF-FIT

Before beginning interpretation of the results, it is useful to examine the goodness-of-fit. This enables one to determine how well the data is represented by the various results and provides a measure of confidence.

6.1 MULTIDIMENSIONAL SCALING

By examining a series of Sheppard diagrams, it is possible to obtain some idea of the goodness-of-fit for multidimensional scaling configurations. Sheppard diagrams for the 2-dimensional, 3-dimensional, 9-dimensional, and 52-dimensional configurations from the Free Sorting Experiment are shown in Figure 6-1 (Sheppard diagrams from the Distributed Experiment are given in Appendix B). The first thing to look for is data that is *clumped* which would indicate degeneracy. Since the data does not appear clumped, it is possible to rule out any effects of degeneracy. Next we look at whether the function is monotonic. This is necessary for a non-metric solution to be considered to have a good fit. Figure 6-1 demonstrates monotonic functions for all dimensions so the solutions are considered valid. An observation is that as the number of dimensions increases, the scatter of the data reduces. This is closely related to the sstress vs. dimension plots of Figure 7-1 which indicates an improved representation with increased dimensionality. The interpretation of these two observations is that the data is well represented by high-dimensional configurations, but in the low-dimensional configurations where there is a large scatter the solutions will contain some inaccuracies.

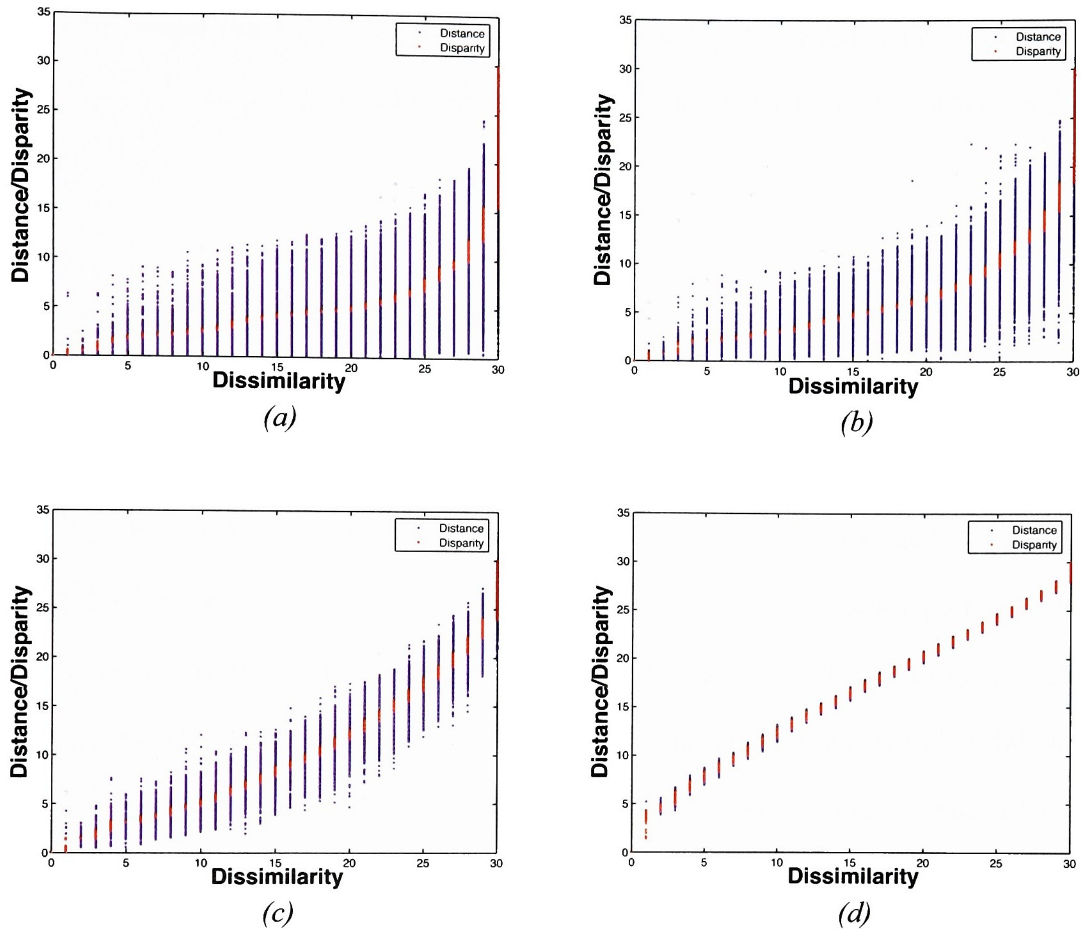


FIGURE 6-1: Sheppard diagrams for the (a) 2-dimensional, (b) 3-dimensional, (c) 9-dimensional, and (d) 52-dimensional multidimensional scaling configurations.

Naturally, we cannot visually examine a 52-dimensional representation of the data so we are limited to evaluating the 2- and 3-dimensional configurations. Figure 6-1 shows us that the goodness-of-fit for these configurations is not very good. This does not mean that we cannot use them at all. Instead, we should simply be aware that the data are not perfectly represented by the configuration and we should expect to see anomalies. In other words, we can look for general trends in the configuration of images but there will be some images that will not fit these trends. While it is possible to use these results to draw general conclusions, we must be careful that we do not interpret too much from the finer details.

6.2 HIERARCHICAL CLUSTER ANALYSIS

The cophenetic correlation coefficient is a logical first step in examining the goodness-of-fit for hierarchical cluster analysis. Table 6-1 lists the cophenetic correlation coefficient for each of the seven linkage methods. Values closer to one indicate a better fit than values closer to zero. We see from Table 6-1 that the Free Sorting Experiment produced very stable solutions using all but the single linkage method. The Distributed Experiment also performed poorly with the single linkage method, but also did not produce solutions that represent the data well for the centroid and median linkage methods. Although the remaining linkage methods indicate reasonable results, it is clear that the Free Sorting Experiment produced cluster trees that better represent the data than the Distributed Experiment.

These results are easily explained. The single linkage method is often susceptible to an effect called *chaining* "which refers to the tendency of the method to cluster together at a relatively low level objects linked by chains of intermediates. ... Because of the chaining effect single linkage may fail to resolve relatively distinct clusters if a small number of intermediate points are present between the clusters." (Everitt, 1974, p. 61) There is evidence that chaining occurred for the single linkage method because numerous levels produced clusters containing only a single image with all remaining images lumped into the last remaining cluster. The centroid and median linkage methods for the Distributed Experiment both produced warnings that the resulting cluster tree was non-monotonic. It is therefore not surprising that the cophenetic correlation coefficient for these methods is poor.

The goodness-of-fit for the Free Sorting Experiment appears to be generally better than that for the Distributed Experiment. Given that the Distributed Experiment was conducted over the internet, that the experiment only produced a sparse matrix, and that an exaggerated amount of noise in the data is expected due to the nature of an internet experiment, the results in Table 6-1 are very reasonable.

TABLE 6-1: Cophenetic correlation coefficient for the seven linkage methods.

	Single	**Complete**	**Average**	**Weighted**	**Centroid**	**Median**	**Ward's**
Free Sorting	0.3817	0.8141	0.8309	0.8394	0.8320	0.8096	0.7872
Distributed	0.2276	0.6334	0.6822	0.5233	0.2892	0.0260	0.6477

Due to the poor cophenetic correlation coefficients, the single linkage results for both experiments will be discarded. The centroid and median linkage methods will also be discarded for both experiments because they produced non-monotonic cluster trees. Although all linkage methods were used during analysis, no further results will be reported for the single, centroid, and median linkage methods.

7. DIMENSIONALITY

Once the goodness-of-fit has been evaluated, the first task was to try and determine how many dimensions are inherent in the data. Multidimensional scaling provides a technique for investigating dimensionality by looking at a graph of stress versus dimension. A detailed explanation of this technique is included in section 3.1.2.

In multidimensional scaling, the quantity called *stress* can be defined as a "measure of the extent to which a geometrical representation falls short of a perfect match." (Johnson and Wichern, 2002, p. 701) As with many statistical quantities, there are many different ways to calculate stress. For the current study, the quantity known as sstress was used. Figure 7-1 shows the *sstress vs. dimension* plot for both experiments. To interpret the results, one must examine the reduction in sstress as a function of the number of dimensions. The value for sstress will necessarily decrease as the number of dimensions increases. Determination of the final dimensionality of the configuration must be made based on principles of interpretability and certain rules-of-thumb. For example, when the value for sstress is plotted against the number of dimensions, there will often be an elbow in the curve. The number of dimensions at which the elbow occurs represents the number of dimensions inherent in the data. Normally, existence of a sharp elbow in the curve will indicate a candidate dimensionality for the data. If no sharp elbow exists, then often there will be a soft elbow which might be used for analysis. Unfortunately, no such elbow can be observed in Figure 7-1. The curve from both experiments decreases so smoothly that no concrete conclusions can be made regarding the dimensionality of the data based on Figure 7-1 alone. However, the complete absence of an elbow might still provide some insight.

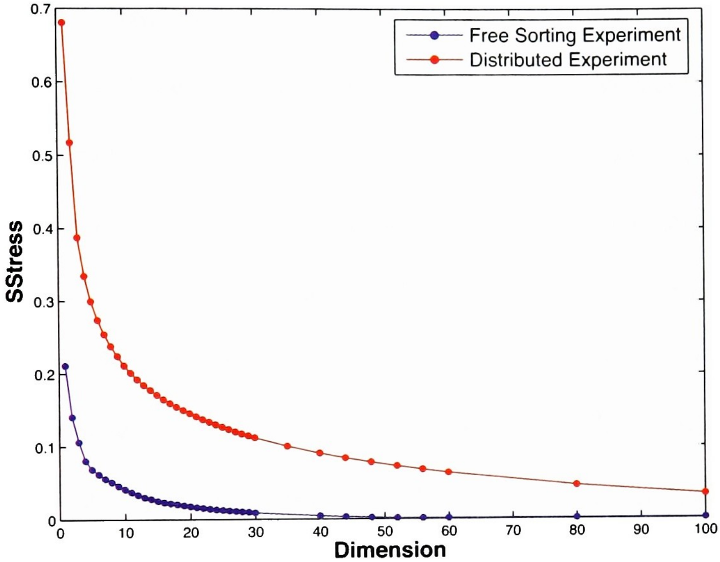


FIGURE 7-1: Dimension vs. SStress plots. There is no clear elbow in either curve to indicate the underlying dimensionality in the data.

It is not surprising that the absolute sstress values are higher for the Distributed Experiment than for the Free Sorting Experiment. Given the nature of the Distributed Experiment, the results are expected to be somewhat less accurate than the results from the Free Sorting Experiment. However, it is noteworthy that the trend that is observed in Figure 7-1 is nearly identical for both experiments.

Another way to explore dimensionality in the data is through cluster analysis. For this task, hierarchical cluster analysis was applied to the data. By examining the silhouette plot for increasing numbers of clusters, it is possible to make some conclusions about dimensionality. Figure 7-2 shows the silhouette plots for 2-5 clusters for the average linkage method (Free Sorting Experiment) which has a result typical of most of the linkage methods (silhouette plots from the remaining linkage methods are found in Appendix C). One interpretation of Figure 7-2 indicates that there is something significant occurring at either two or three dimensions. On one hand, the two-cluster silhouette plot exhibits fairly good unity for both clusters with only a handful of objects having negative values and both clusters showing fairly high silhouette values. On the other hand, the three-cluster silhouette plot shows more objects with negative values but better separation between the clusters (i.e. steeper slopes). The mean value for both groupings is approximately the same, indicating that a valid case could be made for either the two- or three-cluster groupings. Similar findings are evident in the silhouette plots resulting from many of the other linkage methods.

A third method for examining dimensionality is to examine the dendrograms of the cluster trees resulting from hierarchical cluster analysis. The dendrogram is normally used to identify natural cluster divisions in the data. Although this is not exactly the same as dimensionality, if one is able to determine the optimum number of interpretable clusters divisions then this may provide some clues to the true dimensionality. Figures 7-3 and 7-4 show the dendrograms for the Free Sorting Experiment and the Distributed Experiment respectively.

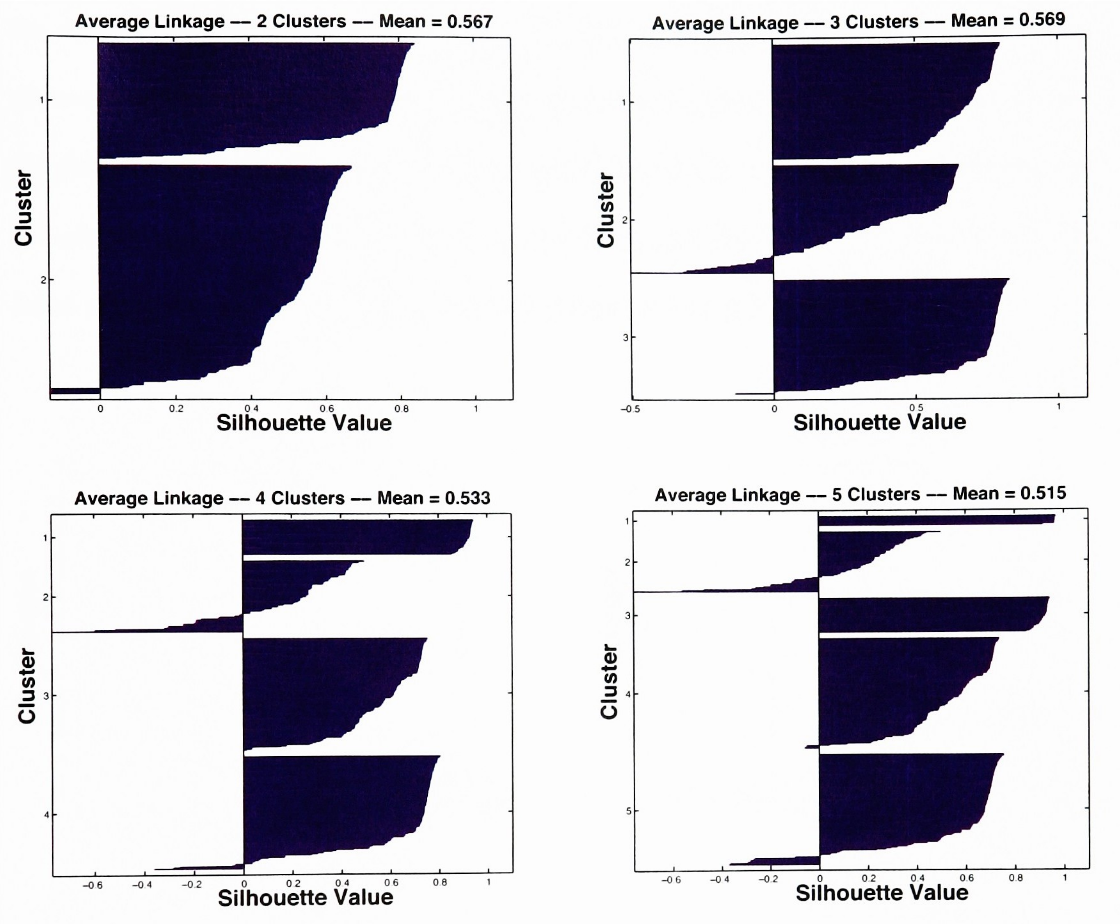


FIGURE 7-2: Silhouette plots for 2-5 clusters of the Average Linkage method from the Free Sorting Experiment.

As we have seen with the other analyses of dimensionality, there is not a clear interpretation of where the natural cluster breaks occur. Figure 7-3 shows cluster divisions that suggest three clusters for each linkage method. A case could easily be made to make the divisions at a lower level such that four, five, or more clusters are formed. Ideally there would be a certain level at which the distances are clearly much further away than the links below, but just as with the SStress vs. Dimension plot (Figure 7-1), the distances gradually taper toward shorter distances. Figure 7-4 displays slightly better division between the clusters suggesting between 4-6 cluster divisions are optimal. Given the results from both experiments, the interpretation of the dendrograms is that between 3-6 cluster divisions provide the most natural cluster divisions. It is important to keep in mind that these divisions are interpretations and other interpretations are possible.

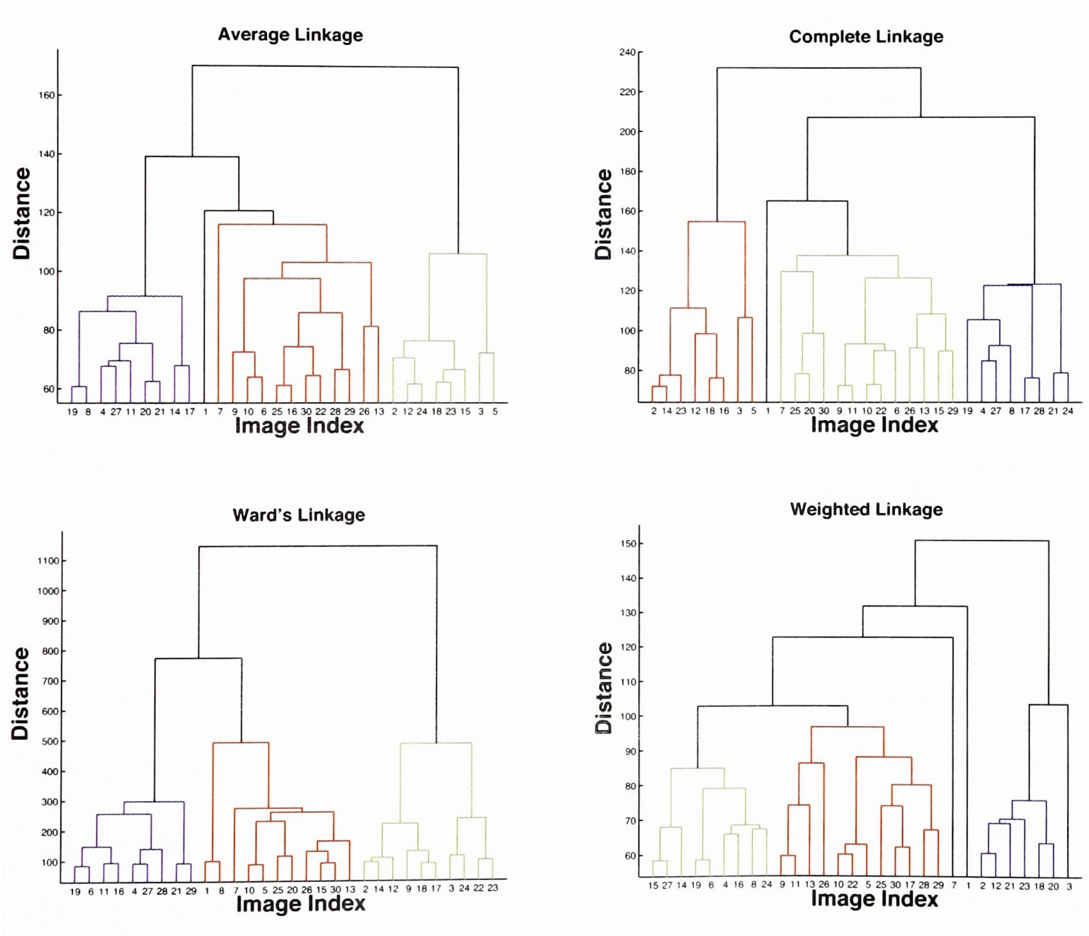


FIGURE 7-3: Dendrograms from the Free Sorting Experiment.

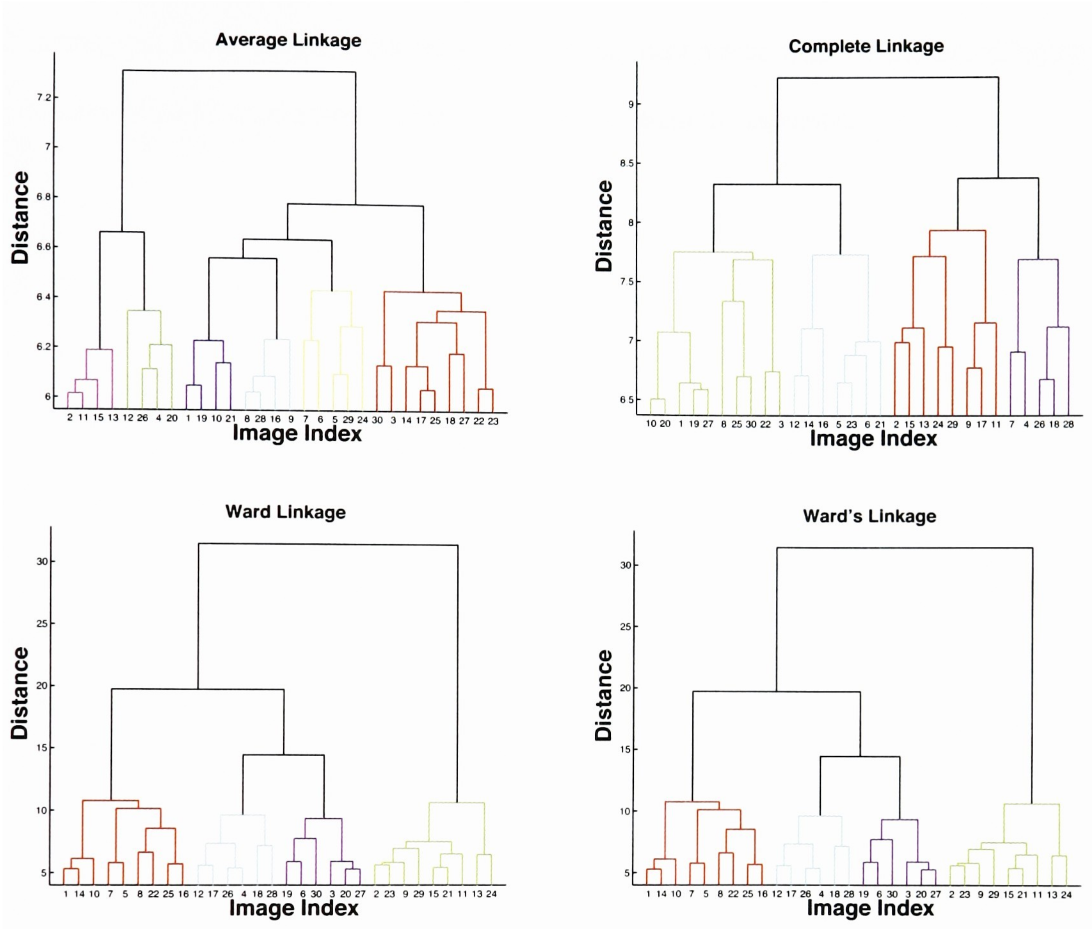


FIGURE 7-4: Dendrograms from the Distributed Experiment.

However, we know intuitively that there cannot be only two, three, or even six fundamental image categories. We also have evidence from the MDS analysis that the dimensionality may be inconclusive. Although it would be best to first identify the dimensionality so that we have some clues about how many categories to seek, we are unable to do this definitively. Perhaps the best interpretation of fundamental dimensionality is that there is no fundamental dimensionality. Instead, what is fundamental about image semantics is their hierarchical nature where each new category division will add some amount of new information about the semantics.

8. CLUSTER INTERPRETATION

8.1 2-DIMENSIONAL MULTIDIMENSIONAL SCALING

By looking at the configuration of the stimuli at solutions with a small number of dimensions, one can interpret the results by looking for trends or finding clusters of stimuli within the configuration that represent similar items. This type of neighborhood interpretation can lead to the discovery of the underlying image categories. Even if we could first determine the true dimensionality of the data, because the primary output of multidimensional scaling is a graphical representation we are still limited to evaluating it using low-dimensional representations—even when the data is determined to be of a higher dimensionality. Therefore the two- and three-dimensional configurations are the only ones that can be easily visually interpreted.

The two-dimensional configuration from the Free Sorting Experiment is shown in Figure 8-1. Although it is impossible to view all images simultaneously due to overlapping, it is nevertheless easy to identify certain trends (careful analysis of all obscured images did not reveal any deviations from the trends that are observable in Figure 8-1). The most obvious trend can be seen by looking at images along dimension 1. Images to the right all have people in them while images to the left have no people. This trend holds very well for the obscured images as well. It is reasonable then to identify our first category as *Images with People* and our second category as *Images without People.* It is important to note that unlike with principle component analysis—where the first dimension represents the most important information, the second dimension represents the next most important information, etc.—no conclusions can be made about the importance of the order of the dimensions. Additionally, recall that the orientation of the configuration is arbitrary and only the inter-relationship between the samples has significance. In other words, we cannot conclude from interpretation of multidimensional scaling that *Images with People* and *Images without People* are the most important dimensions nor are we restricted to interpreting along the coordinate axes, though in this case there appears to be good correlation with the coordinate axes.

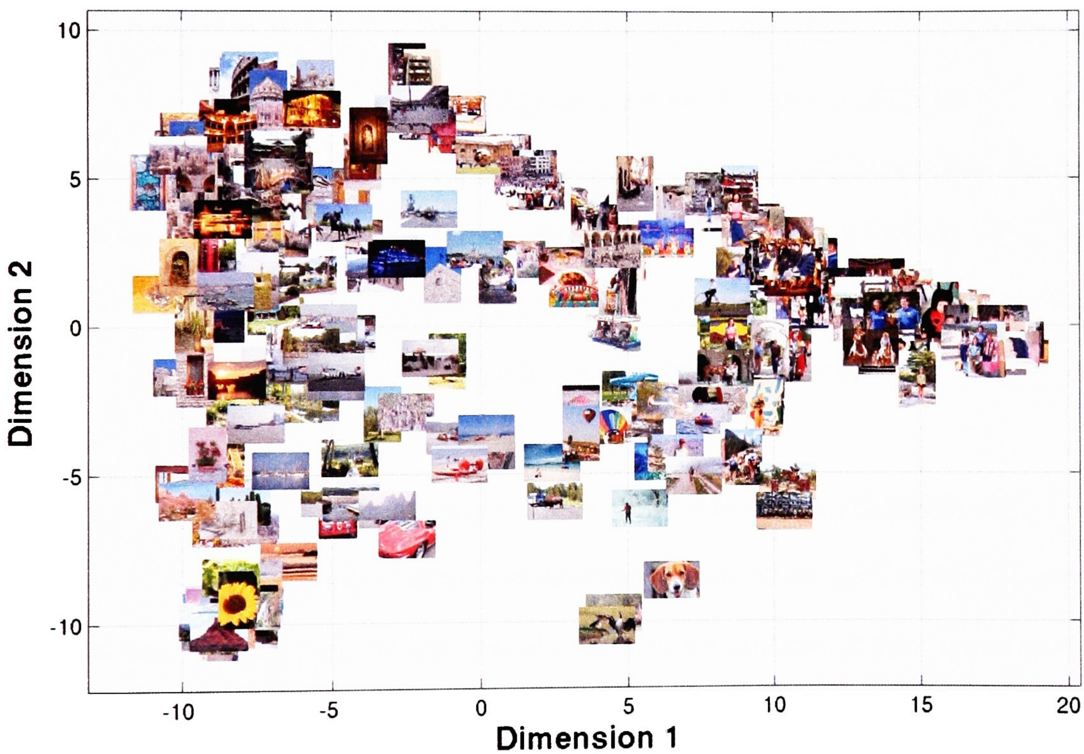


FIGURE 8-1: Free Sorting Experiment multidimensional scaling 2-dimensional configuration.

Further examination of Figure 8-1 reveals another trend. In general, the majority of images with negative values for dimension 2 (images near the bottom of the figure) include natural scenes, landscapes, etc., while the majority of images with positive values for dimension 2 (images near the top of the figure) display varying degrees of man-made objects such as tables, sculptures, buildings, etc. A transition from natural images to images featuring man-made objects is clearly the primary characteristic of dimension 2. This observation is emphasized by considering the first two categories. Images at the extreme of dimension 1 appear to be primarily images with people filling the frame without much reference to their surrounding environment. As one moves toward the other end of the dimension, it can be seen that images of people in natural/rural scenes appear in the lower portion of the configuration and images with people in city scenes (i.e. man-made environment) seem to occupy the upper portion of the configuration. Therefore, our third category is *Natural Images* and our fourth category is *Images with Man-Made Objects*.

The two-dimensional configuration from the Distributed Experiment is shown in Figure 8-2. It is immediately obvious that the overall shape of the configuration is very different from the one obtained from the Free Sorting Experiment. In three dimensions, the circular shape is expanded into a spherical configuration. The same circular configuration was observed in a previous study (Rogowitz, et al., 1998). It is unclear exactly why the computer experiments are producing such geometrically symmetric results, but it is very likely nothing more than an artifact of the experimental design. This seems particularly likely since it occurred identically in both separate studies. Both computer experiments produced a sparse data set which might account for the circular/spherical configuration. In contrast, the tabletop sorting experiment in (Rogowitz, et al, 1998) also produced a circular configuration. Rogowitz speculated that this was perhaps due to the use of metric multidimensional scaling and that using non-metric multidimensional scaling instead might produce a more natural configuration. In the present study, non-metric multidimensional scaling was used and the configuration is indeed non-circular.

If the circular shape is ignored, it becomes clear that the same trends observed in Figure 8-1 are also present in Figure 8-2. Not only are the same trends observed, but a detailed examination of individual images in both figures reveals that most of the images are located in similar regions of the plot and are similarly oriented with respect to nearby images. This provides good initial confirmation that results of both Experiment I and Experiment II are reliable and that both experiments have measured the same psychological process.

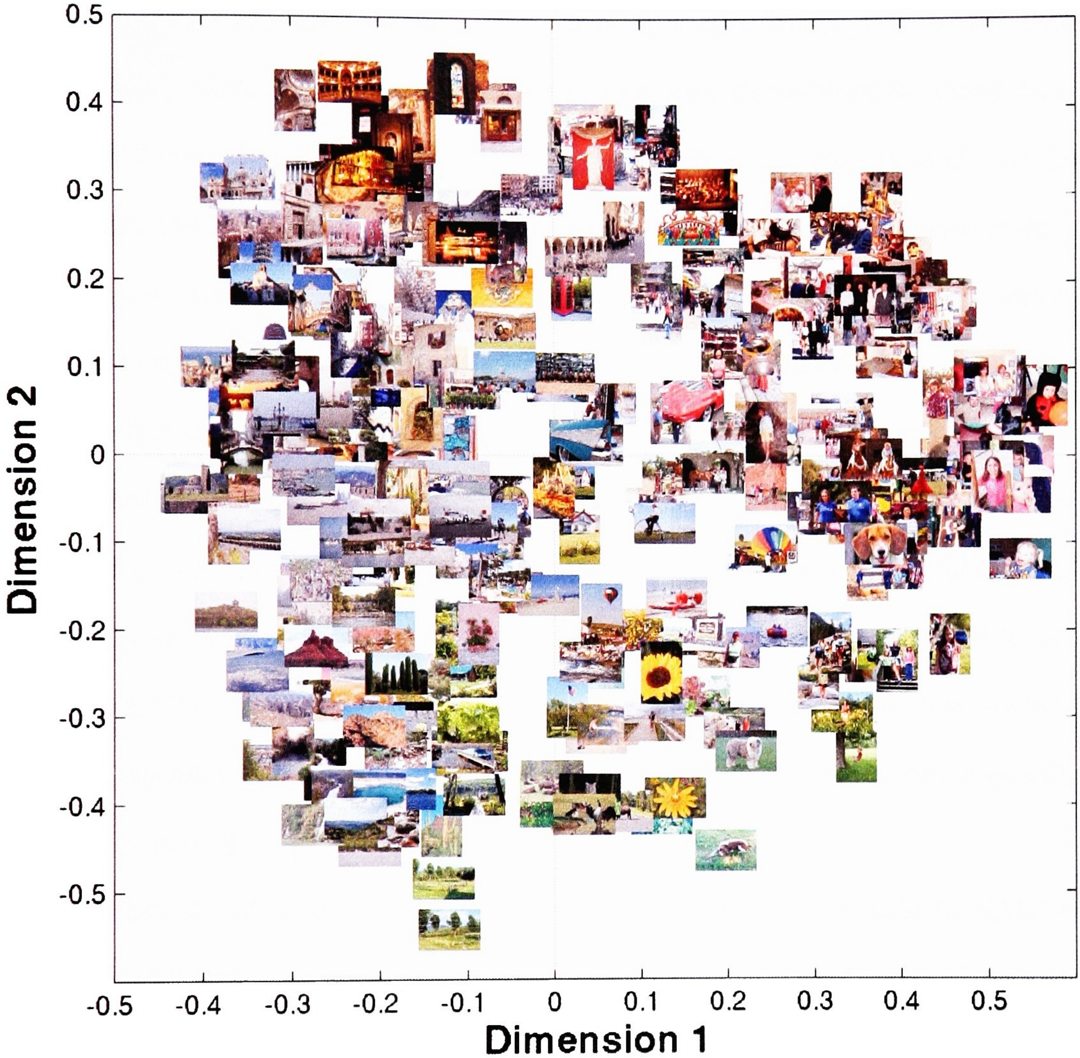


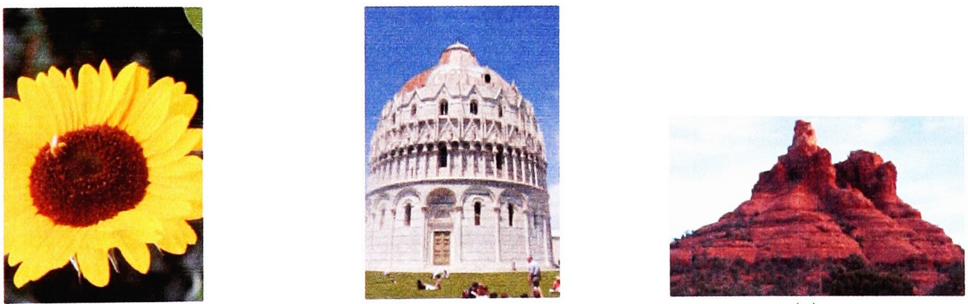
FIGURE 8-2: Distributed Experiment multidimensional scaling 2-dimensional configuration.

8.2 3-DIMENSIONAL MULTIDIMENSIONAL SCALING

Looking at the 3-dimensional configuration is a little more difficult. Because the three-dimensional configuration is impossible to view on a printed page in three dimensions, it is presented in slices with each slice representing 10% of the third dimension. By viewing these in sequence, one can get some idea of the three-dimensional nature of the configuration. Figure D-1 in Appendix D shows the ten slices for the Sorting Experiment and Figure D-2 shows the ten slices for the Distributed Experiment.

Before continuing with the interpretation of the third dimension, it is necessary to address the issue of orientation. Recall that the dimensions of a multidimensional scaling configuration are not orthogonal nor is the orientation of the configuration necessarily aligned with the coordinate axes. Evidence of this can be seen in Figure D-1 and Figure D-2. Notice how the second dimension has reversed its orientation with respect to the two-dimensional configuration (Figure 8-1). Although the first two dimensions were well aligned with the coordinate axes in the two-dimensional configuration, this may not hold true with the third dimension. In fact, it is possible that a valid dimension cuts diagonally through the three dimensions making it difficult to identify.

While not as obvious as the first two dimensions, there is a definite trend that is discernable along the third dimension and it also happens to align fairly well with the coordinate axis. A relationship can be observed between an image's position along the third dimension and the *perceived proximity* of the primary subject of the photograph. *Perceived proximity* is defined as the distance from the observer's perceived position in the scene to the focal point of the image. In other words, it can be thought of as the degree to which a subject fills the frame relative to its physical size. For example, the three images in Figure 8-3 all fill approximately the same amount of area in the image frame, but one has the perception that one is physically much closer to the flower than to the building or mountain. This is because we know that a flower is much smaller than the other objects and therefore the distance to the flower must be much shorter in order to achieve the relative perspectives. It is also assumed that the observer has some experience and knowledge of the effect of the focal length of a camera lens. In other words, to achieve the three images in Figure 8-3, lenses of different focal lengths are required and this is intuitively taken into account by the observer. The physical proximity of the photographer may be different than the perceived proximity of the observer due to a reduction or enlargement of the image within the camera system. It is the perceived proximity that is relevant to the image semantics.



(a) *(b)* *(c)*

FIGURE 8-3: Examples of perceived proximity. The observer has the perception of being physically closer to the flower (a) than the building (b) or the mountain (c) even though they all fill approximately the same amount of the image frame.

To explore the idea of perceived proximity, it will help to identify some examples. The observed trend can be stated as the transition from a perceived proximity of very close at one end of the dimension to a perceived proximity of very far at the opposite end. This appears to hold true for images that are similar, however the position along the third dimension does not remain constant for images of different categories. Nevertheless, the trend appears to be stable. Figure 8-4 demonstrates this transition using images taken from the same region (based on the first two dimensions) of the three dimensional configuration from the Sorting Experiment. Each image is labeled with the region of the third dimension that the image occupies. Notice how at the lower end of the third dimension we see only a single flower. As we move along the dimension we can see an entire flower bed. Move further still and we begin to see a greater part of the surrounding scene. Continue along the dimension and we begin to see a full landscape.

(a) *(b)* *(c)* *(d)*

FIGURE 8-4: Example of perceived proximity from the 3-dimensional configuration of the Sorting Experiment. Images along the third dimension are found at (a) 20-30%, (b) 30-40%, (c) 40-50%, and (d) 50-60%.

This trend of beginning with a very close perceived proximity and slowly moving farther away can be observed in numerous regions of the same configuration. Figures 8-5 and 8-6 show two more examples of this trend. Although not all images conform this observation, the majority of images do appear to conform and the evidence for it is too strong to ignore.

FIGURE 8-5: Example of perceived proximity from the 3-dimensional configuration of the Sorting Experiment. Images along the third dimension are found at (a) 10-20%, (b) 20-30%, (c) 30-40%, (d) 40-50%, (e) 50-60%, and (f) 60-70%.

FIGURE 8-6: Example of perceived proximity from the 3-dimensional configuration of the Sorting Experiment. Images along the third dimension are found at (a) 30-40%, (b) 40-50%, (c) 50-60%, and (d) 60-70%.

Note that the images in Figure 8-4, Figure 8-5, and Figure 8-6 do not remain in the exact same region with respect to the first two dimensions. That is, as one moves up along the third dimension, the coordinates for the other two dimensions tend to drift. This was explained earlier by the fact that a dimensional axis can have any orientation and in this case, the axis is not in complete alignment with the coordinate axis.

8.3 DUAL SCALING

Although it is not feasible with multidimensional scaling to examine high-dimensional configurations, the higher dimensions can be analyzed using dual scaling. By looking at the objects at the extremes of the various dimensions, some characteristics about that dimension may be inferred. To facilitate this, the first and last eight images from each dimension were explored. Figure 8-7 shows the extremes of the first two dimensions from the Free Sorting Experiment. The top eight images represent one extreme and the bottom eight images represent the opposite extreme. The extremes for dimensions 3-10 of the Free Sorting Experiment and dimensions 1-10 of the Distributed Experiment are found in Appendix E. Although the first 50 dimensions were examined, due to space limitations only the first 10 are included in this thesis.

(a) *(b)*

FIGURE 8-7: Images at the extremes of (a) dimension 1 and (b) dimension 2 from dual scaling analysis for the Free Sorting Experiment.

Figure 8-7 (a) shows a very clear division between the images at the extremes. At one end there are images containing rocky landscapes that include mountains. At the other end are images of people. More specifically, there are primarily close-up images of young children. Figure 8-7 (b) has interior images of ornate, historical buildings which includes the sculpture/art that is often found in these locations. The opposing extreme contains all images of animals. Analysis continued in this manner through the first 50 dimensions and the results were compiled. When a group of images at the extreme of a dimension had an interpretable characteristic, that characteristic was considered one of the potential categories inherent in the underlying data structure.

Some categories that were identified were related to categories from other dimensions but revealed a finer resolution of detail. For example in one dimension there was a mixture of different types of images with people, but another dimension exhibited only close-up images of children. In such cases, the categories are combined and only those at the finest resolution are reported. Due to this condensation of categories, it is no longer possible to report them in any particular order with respect to the dual scaling dimensions. The list of categories resulting from this analysis is given in Table 8-1.

TABLE 8-1: Categories resulting from dual scaling analysis of the images at the extremes of the dimensions.

Dual Scaling Categories
people (posing)
people (candid – inactive)
people (candid – active)
children (close-up)
animals
floral
rocky landscapes / mountains
waterscapes
water sports
cityscapes w/water
bridges
modern architecture
food / dining
vacation
night

8.4 HIERARCHICAL CLUSTER ANALYSIS

Data from both the Free Sorting Experiment and the Distributed Experiment were divided into clusters using the complete, average, weighted, and Ward's linkage methods. This was repeated for 2-20 fixed clusters for each of the four linkage methods. The images that belonged to each resulting cluster were examined and prominent identifying characteristics, if any, were compiled.

For example, Figure F-1 in Appendix F shows the image clusters resulting from the 5-cluster division of the Free Sorting Experiment using the average linkage method. Cluster 1 is clearly all images of animals. It would not be appropriate to identify this as *animals outdoors* because all but two images in the experiment that have animals are images of animals outdoors. Cluster 5 contains all images with people. Note that with few exceptions, there are no images in the other categories that have people at all, the exception being city scenes that show a crowd of people from a distance. Cluster 3 appears to be composed of images of natural scenes such as landscapes, waterscapes, and mountains. The remaining two clusters are not quite as easy to interpret. Cluster 2 appears to also have images that are primarily of natural scenes. However, unlike cluster 3, there is a definite presence of human influence. For example, Mount Rushmore is certainly an image of a mountain, but the four faces indicate a clear human influence. Also notice how the waterscapes of cluster 2 all contain either boats or buildings along the coastline while the waterscapes of cluster 3 are purely natural. If we relate this to the multidimensional scaling configuration in Figure 8-1 we will see that the images of cluster 3 will be found towards the bottom of the second dimension while the images of cluster 2 will be found somewhere between the top (images with man-made objects) and bottom (natural images). Finally, cluster 4 seems to have primarily images of man-made objects. In particular, the images are mostly of architecture. Although this cluster has the weakest cohesion of the five, as more cluster divisions are introduced and the clusters become more refined, the divisions become more cohesive.

Figure F-2 shows the first two image clusters resulting from the 6-cluster division of the Free Sorting Experiment using the average linkage method. The remaining clusters were unchanged from the 5-cluster groupings. Notice how in the previous stage all images with people were clumped together into a single grouping (cluster 5) and now they have divided into two separate clusters. Close examination of the two clusters reveals that in cluster 1 all of the people appear to be posing for the camera or are aware of the picture taking process in the majority of the images. On the other hand, cluster 2 images have a tendency to be candid photographs. Thus we have an identifiable division of a previous cluster. Later divisions will reveal the candid photographs dividing into one group where the people are inactive (i.e. sitting, standing) and another group where the people are engaged in some activity (i.e. swimming, riding horses). This is an example of the hierarchical nature of the method and of the hierarchical structure of image semantics.

Interpretation of the cluster divisions continued until the cluster divisions no longer made logical sense and no identifiable characteristic between the clusters could be determined. Cluster divisions for each linkage method for both experiments were cross-referenced and those that appeared in a majority of the linkage methods were considered stable and were accepted. However, some cluster divisions either appeared in only one method or were weakly identifiable and were therefore discarded.

Note that when more than one linkage method produced the same cluster divisions, such as candid and posed images of people branching out from a single cluster of all people, this usually did not occur at the same level. For example, the 4-cluster divisions of the complete linkage contained a single grouping of *people* images, but the 5-cluster groupings included the posing and candid branches of the people category. This occurred at a different level than with the average linkage method. What is important is not the level at which it occurred, but that the divisions occurred in most linkage methods and followed the same trends.

All identifiable image groupings were compiled into a hierarchical tree (Figure 8-8). Although Figure 8-8 looks somewhat like a dendrogram, it is important to understand that it is not a dendrogram. Figure 8-8 is a summary of the cluster interpretations described above. It is presented in a tree-like structure to emphasize the parent/child relationship between the various branches in the hierarchy. There were four main groupings from which all others branched out—people, natural, manmade, and a fourth grouping containing a non-descript combination of images of both natural and manmade scenes (indicated by the small square). These categories, along with the categories identified in the dual scaling results (Table 8-1), will be used as an aid in interpreting possible characteristics in the local multidimensional scaling configurations.

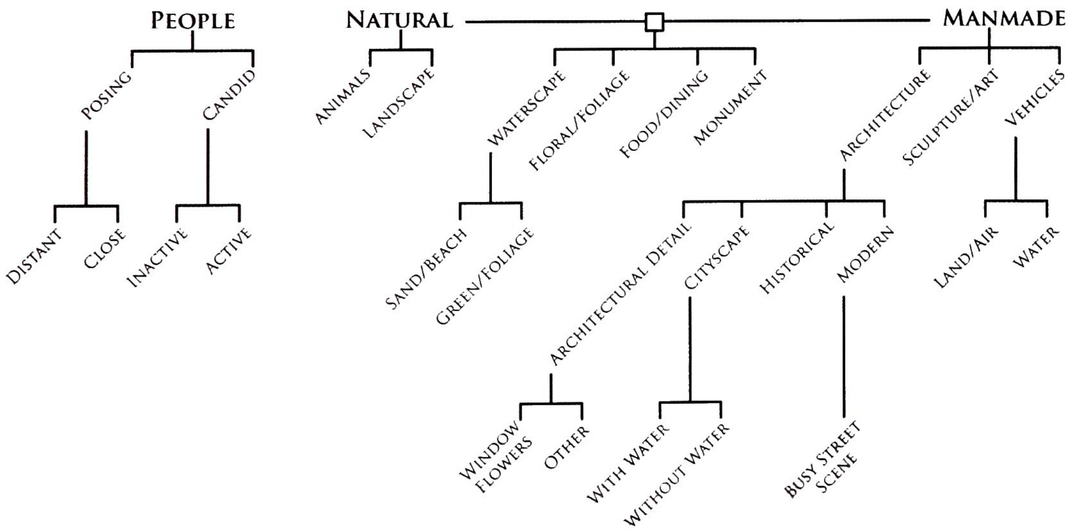


FIGURE 8-8: Cluster interpretations from hierarchical cluster analysis.

8.5 LOCAL MULTIDIMENSIONAL SCALING CONFIGURATIONS

Sometimes it is possible to detect additional structure in data by re-applying multidimensional scaling to data from local regions of the configuration. This process will often produce configurations that make is easy to identify characteristics in the data that were obscured in the original configuration. There are several ways to determine which objects to include in the local processing. In one method, the global configuration can be divided into a regular grid and all objects falling within a particular division are selected. In another method, clusters of points from the global configuration can be selected based on visual interpretation.

For this study, clusters from the hierarchical cluster analysis were utilized to help determine the regions for local multidimensional scaling. To determine which cluster divisions to select, the global configuration was plotted with the points colored based on the cluster divisions. Figure 8-9 shows an example of this type of plot for the 6-cluster division of the weighted linkage method for the Sorting Experiment. Plots similar to Figure 8-9 were made for 3-20 clusters for each of the 4 linkage methods and repeated for each of the two experiments. Each plot was examined to see which had the maximum number of clusters with the minimum amount of spatial overlap between clusters. The 6-cluster weighted linkage clusters seemed to have a good balance of these two properties. For the Distributed Experiment, the 6-cluster Ward's linkage clusters were selected (Figure 8-10). The regions were not selected based on any strict rules. Instead the decision of which cluster arrangement to select was based on interpretation and practicality.

Next the centroid for each of the six clusters was calculated for the configuration. Each centroid was used as the center of a local region and points were selected based on a fixed radius from the center. For the Sorting Experiment, all points within a radius of 5 were included and for the Distributed Experiment, all points within a radius of 0.17 were included (Figures 8-11 and 8-12 respectively). Finally, the selection of local clusters of points was processed with multidimensional scaling using the same procedure for the global configuration described in section 5.1.

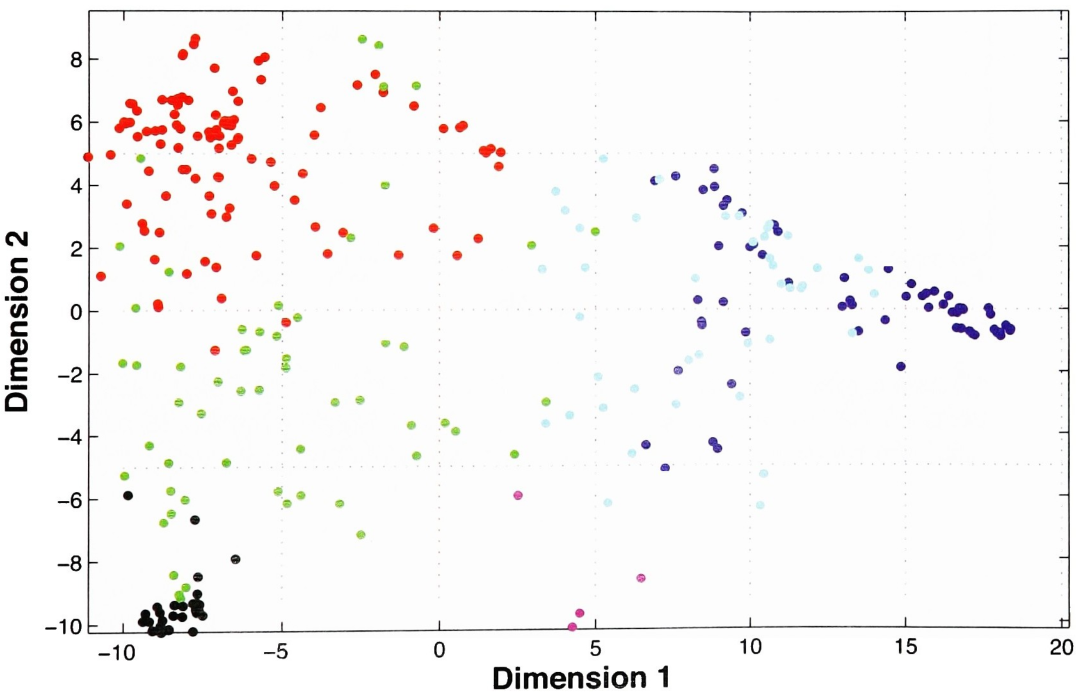


FIGURE 8-9: Multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment. Images are colored based on the 6-cluster grouping using the weighted linkage method of hierarchical cluster analysis.

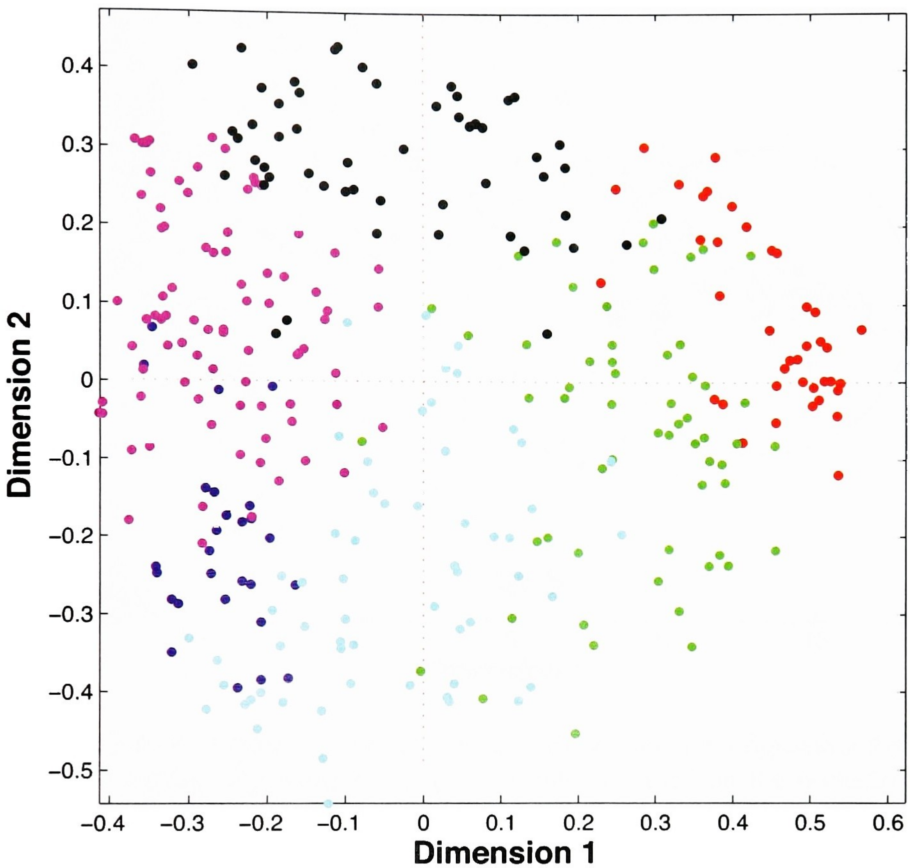


FIGURE 8-10: Multidimensional scaling 2-dimensional configuration for the Distributed Experiment. Images are colored based on the 6-cluster grouping using the Ward's linkage method of hierarchical cluster analysis.

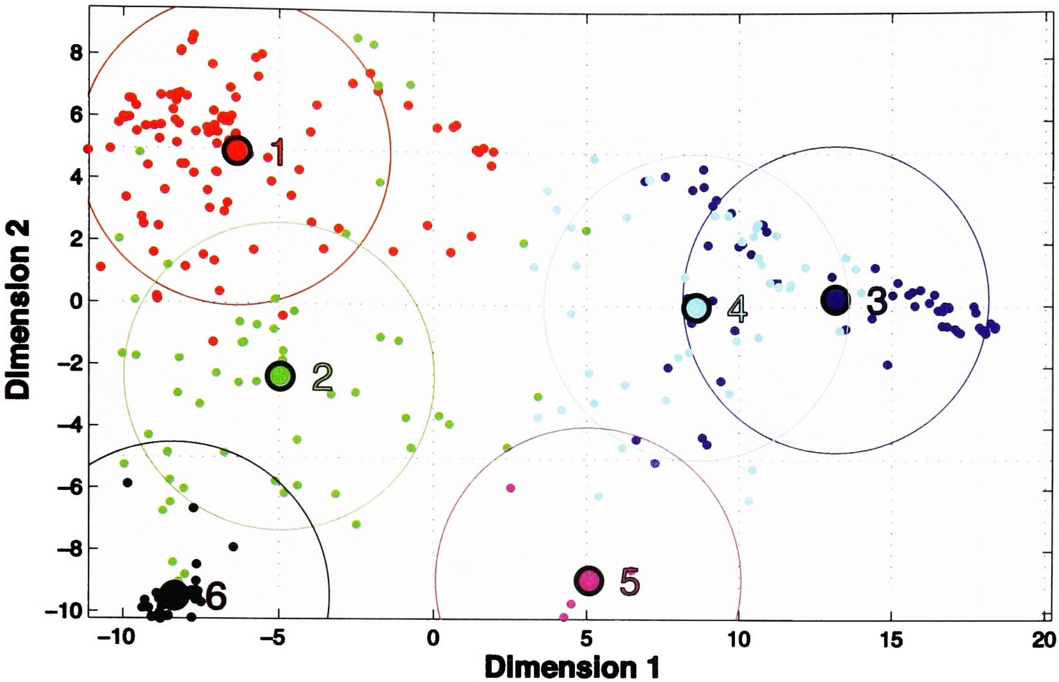


FIGURE 8-11: Multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment. Images are colored based on the 6-cluster grouping using the weighted linkage method of hierarchical cluster analysis. Local multidimensional scaling was applied to image groupings within each circle. The circle centers are the centroids from the corresponding hierarchical cluster analysis clusters.

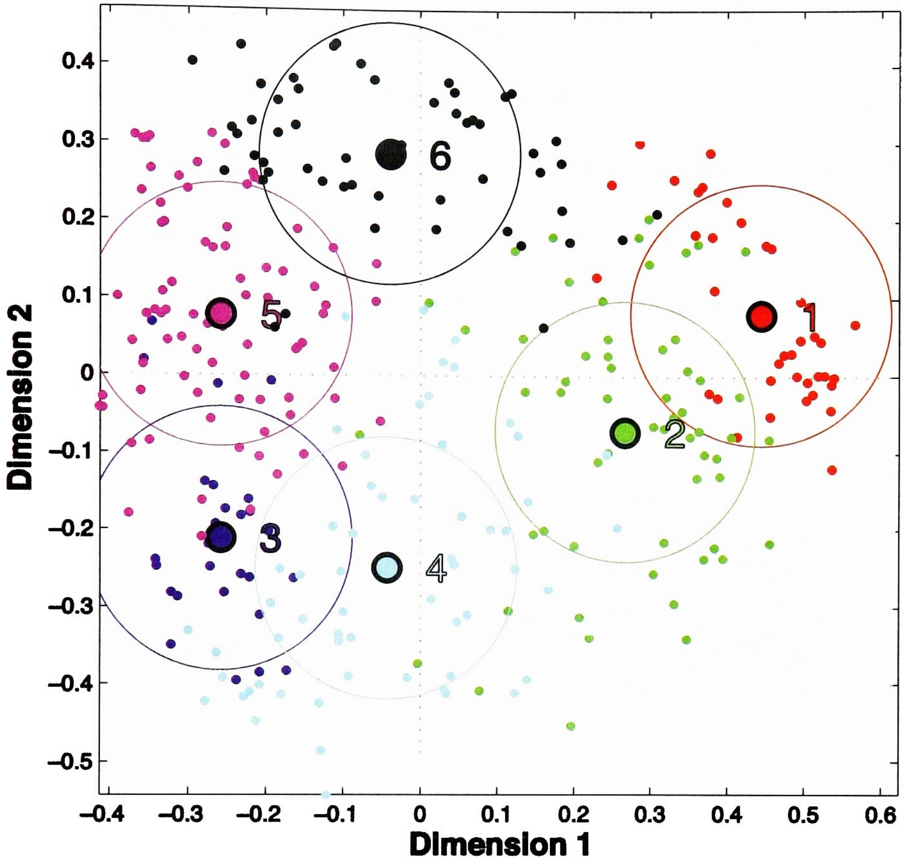


FIGURE 8-12: Multidimensional scaling 2-dimensional configuration for the Distributed Experiment. Images are colored based on the 6-cluster grouping using the Ward's linkage method of hierarchical cluster analysis. Local multidimensional scaling was applied to image groupings within each circle. The circle centers are the centroids from the corresponding hierarchical cluster analysis clusters.

Interpretation for the local configurations proceeded in the same manner as for the global configurations. Figures 8-13 and 8-14 show two of the local configurations from the Sorting Experiment. The remaining local configurations from both experiments are found in Appendix G. Figure 8-13 shows the local configuration for cluster 2. The images near the bottom all have foliage and floral themes. This characteristic was already evident in the global structure. However, notice how all of the images near the top-right are waterscapes or have water themes. This is a new characteristic that was previously obscured in the global configuration and provides good support for the various water-related categories identified in dual scaling and hierarchical cluster analysis.

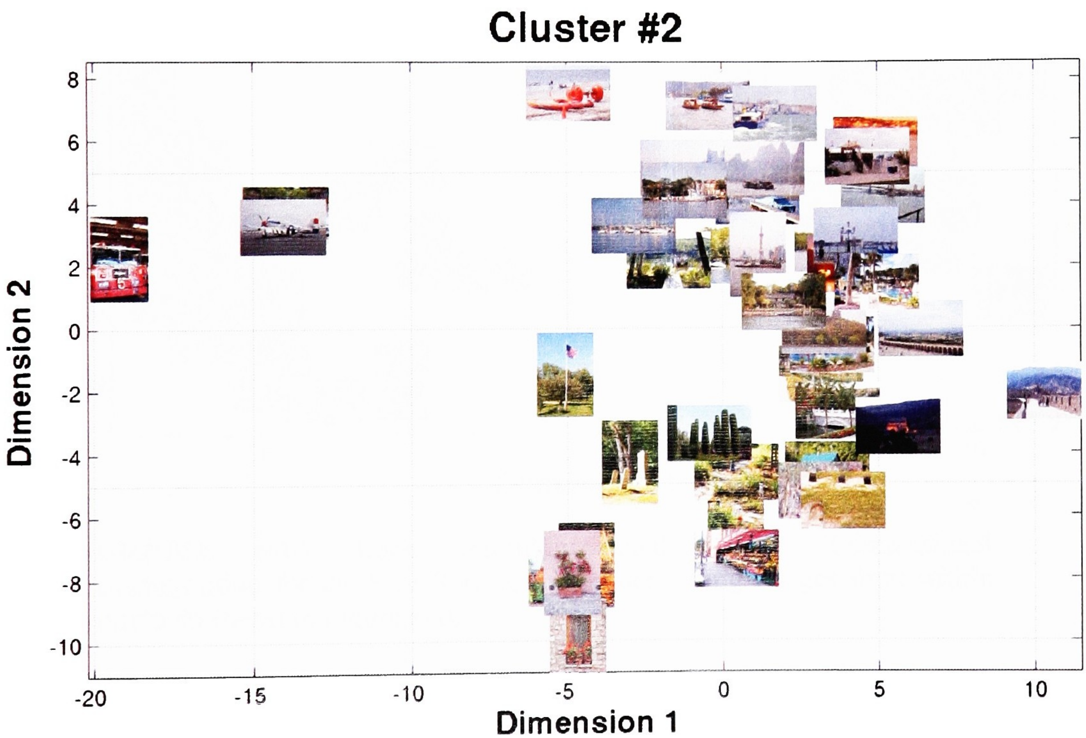


FIGURE 8-13: Local multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment using images from within circle #2 (green) in Figure 8-10.

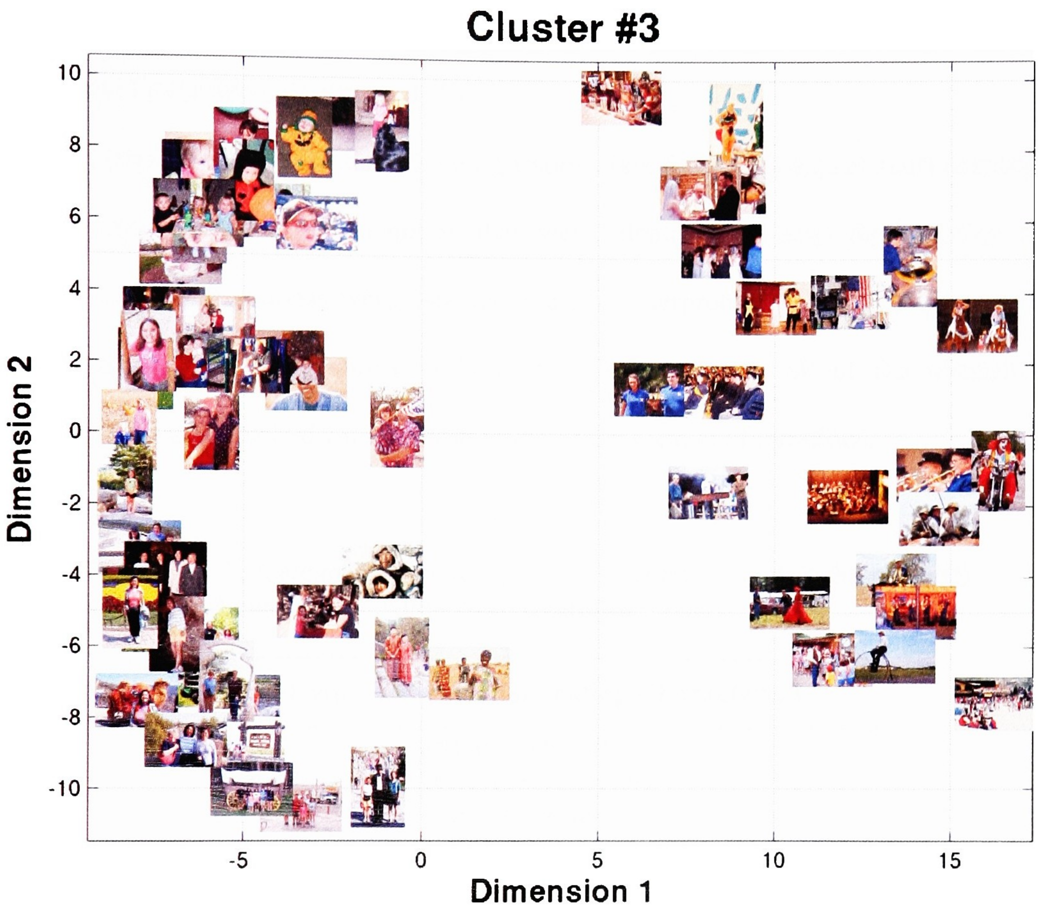


FIGURE 8-14: Local multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment using images from within circle #3 (blue) in Figure 8-10.

Figure 8-14 also reveals some new characteristics that were not evident in the global configuration. The most striking feature is the clear division down the center of the configuration. Although the global configuration clearly showed the presence of people as an important characteristic, it was not clear if there were further divisions. In Figure 8-14 we see that all of the images on the left side are images of people posing for the camera and the images on the right side consist of candid photographs of people. This is very good confirmation of this characteristic which was also detected in both dual scaling and hierarchical cluster analysis.

Interpretation of all 12 local configurations (six from each experiment) continued in the same manner. The categories that were identified through dual scaling and hierarchical cluster analysis were used as an aid in interpreting the local configurations by suggesting certain groupings to look for. Table 8-2 lists all of the identifiable characteristics that resulted from interpretation of the local configurations.

TABLE 8-2: Categories resulting from interpretation of 2-dimensional local multidimensional scaling configurations.

Local Multidimensional Scaling Characteristics
modern architecture
historical architecture
art / sculpture
bridges
waterscapes
floral
foliage
people (posing)
people (candid – inactive)
people (candid – active)
children
animals
natural landscapes
rocky landscapes
water
natural images with human influence

8.6 FINAL FUNDAMENTAL CATEGORIES

By compiling the identifiable categories resulting from all thee methods of analysis — global and local configurations of multidimensional scaling, dual scaling, and hierarchical cluster analysis — a new list of the final fundamental categories is formed (Table 8-3). These 34 categories are the ones that have proven to be stable across at least two of the methods of analysis.

TABLE 8-3: Final list of fundamental categories resulting from both experiments and all methods of analysis.

FINAL FUNDAMENTAL CATEGORIES
people
people – posing
people – posing (distant)
people – posing (close)
people – candid
people – candid (inactive)
people – candid (active)
natural
animals
landscape
manmade
sculpture/art
vehicles
vehicles – land/air
vehicles – water
architecture
cityscape
cityscape w/water
cityscape w/out water
architectural detail
architectural detail – window flowers
historical architecture
modern architecture
perceived proximity
waterscape
waterscape sand/beach
waterscape green/foliage
floral/foliage
food/dining
bridges
landscapes – rocky/mountainous
landscapes – green
day
night

9. DISCUSSION

Having identified a set of fundamental categories, one may be tempted to conclude the study here. However, it may be worthwhile to consider the results further. Perhaps there is a deeper understanding that can be gained by considering the composition of the categories themselves. Before continuing, it is helpful to briefly review a different human perception — *color.*

9.1 COLOR PERCEPTION

It is easy to divide the full range of perceived colors into a set of distinct categories. Among other important contributions, Berlin and Kay (1969) discovered extraordinary similarities in color vocabulary across 20 languages and introduced the concept of *basic color terms* of which there are 11 in English (*black, white, red, green, yellow, blue, brown, pink, orange, purple, and gray*). The National Bureau of Standards introduced the *ISCC-NBS color names dictionary and the universal color language* which defined color names for 267 regions in Munsell color space using a combination of basic hue names (*red, orange, yellow, green, etc.*) and a variety of modifiers including as *dark, medium, light, grayish, vivid, brilliant, and pale* (Mojsilovic, 2002). In the first case we can think of dividing all colors into 11 categories. In the second case we are dividing all colors into 267 categories.

While the first color system divides colors into broad, fundamental categories, the second defines a finer resolution which allows a greater degree of discrimination. Imagine if systems of color categories such as these were the only way in which we defined color. It would extremely difficult, if not impossible, to design and engineer the multitude of color technologies that exist today if we were restricted to using a finite set of categories to describe color. Picture the challenge of developing color broadcast television using only categorical color descriptions!

There are better ways to model color than mere categorical descriptions. Color appearance only requires five perceptual dimensions (*lightness, chroma, hue, brightness, and colorfulness)* to unambiguously describe the appearance of a color (Fairchild, 2005, Ch. 4). In other words, using these five perceptual dimensions, all 267 categories of the ISCC-NBS color naming system, and many more, can be identified.

9.2 PERCEPTUAL IMAGE SEMANTIC DIMENSIONS

Image semantics are usually modeled using basic categorical descriptions (Cox, et al., 2000; Greisdorf and O'Connor, 2002; Chan, et al., 2006). Sometimes a more comprehensive approach is employed that accounts for a hierarchical structure of the semantics under investigation (Boutell, et al., 2003; Tian, et al., 2005; Rorissa and Hastings, 2004). Nevertheless the descriptions are still categorical by nature. There have been no attempts known to the author to model image semantics according to their underlying perceptual dimensions. Just as color is better modeled using the perceptual dimensions, image semantics could benefit greatly by a similar approach.

In the color naming example, basic color names are used in combination with various modifiers such as *light green* and *dark blue*. Notice how the modifiers of *light* and *dark* are direct correlates of the *lightness* dimension of color appearance. By analyzing the final fundamental categories in Table 8-3, it is possible to construct a perceptual semantic model based on similar principles. For example, there are seven categories that are connected with the presence or absence of water. If water were used as a modifier, similar to light and dark, then we could define a *wetness* dimension that can be used to indicate the presence of water in an image. Similarly, there are seven categories indicating the presence of people. If one disregards the differences between these categories, then the modifiers *with people* and *without people* can indicate another perceptual semantic dimension. In fact, a case can be made to include the *animal* category in this dimension as well. A closer examination of Figure 8-1 reveals that all images of people are at one end of the first dimension and all images without people are at the other end while images with animals appear to be positioned somewhere in between. This is interpreted to mean that the scale is not only representing the presence of people but rather the presence of any living creature where people are considered the most important followed by animals and finally the absence of all living creatures. We can call this the *humanness* dimension to describe the degree to which a living creature is similar to a human.

Following these principles, it is possible to represent the 34 final categories using only 10 perceptual semantic dimensions (Table 9-1). These 10 dimensions represent the underlying perceptions that are responsible for making categorical assignments. In forming the 10 dimensions, a few of the categories, such as *sculpture/art*, *vehicles*, and *food/dining* were omitted because they were not perceived to be fundamental in nature. In other words, they are special cases of other categories and only appeared as independent groupings in this experiment because there happened to be enough images that contained them included in the test images. Had there been at least 10 images of soccer balls, footballs, baseballs, etc. then a new category of *sports balls* would have likely appeared but that would not mean that *sports balls* is a fundamental category. Rather, it is just a special case of manmade objects. Therefore, the interpretation is to omit these categories.

TABLE 9-1: Proposed perceptual image semantic dimensions.

Perceptual Image Semantic Dimensions
Humanness
Artificialness
Perceived Proximity
Candidness
Wetness
Architecture
Terrain
Activeness
Lightness
Relative Age

A multidimensional perceptual semantic space provides a convenient mechanism for describing image semantics. If each dimension were properly scaled (the current study is only suggesting the concept of a semantic space—to fully implement it would require psychophysical experiments to scale each dimension) then similar images would be located within the same sub-region of the space. This accomplishes numerous objectives as outlined in the introduction.

One use of image semantics is to aid in image-dependent processing such as gamut mapping and contrast adjustment. If the descriptors for an image are pre-computed, then it is only necessary to identify the coordinates in semantic space to determine what type of image it is. If they are not pre-computed, then the processing required to classify the image is greatly reduced since it is only necessary to identify 10 image descriptors instead of a nearly endless list of possible categories. Of course, it may be required to calculate several quantities before the image can be rated for any one dimension, but the result will be an image description that is likely more complete than if every category were to be rated individually.

Image retrieval from large databases is another active area of research that utilizes image semantics. By using a multidimensional semantic space rather than categorical descriptions it is possible to create a much more flexible search engine. Because similar images are located in the same sub-region of the space, a search engine would be able to simply pick a semantic *center* and collect all images within a certain radius. The larger the radius is the more broad the search results will be. As the radius decreases, it is easy to converge on a desired result. This structure also solves the problem where images do not belong to a fixed number of semantic categories—images can be belong to many categories simultaneously. It is a nearly impossible task to pre-judge every category that an image might belong to but this is not necessary with a semantic space.

If a particular type of image is determined to be of importance, then it is still possible to provide more detailed information without adding additional dimensions. For example, if a database is to store images of wildlife and it is necessary to distinguish between images of tigers, lions, leopards, antelope, bears, wolves, whales, dolphins, tuna, squid, etc., then performing a scaling experiment on these animals will help determine their coordinates. Recall that animals are defined along the *humanness* dimension. Scaling the above creatures by rating each species according to the degree to which it has human characteristics, the various species will naturally align themselves along this single dimension. Most people will probably see more human qualities in a wolf than in a squid which would place the wolf closer to the human end of the scale. All of the big cats will probably be rated nearly the same, but further scaling can be performed to divide the big cats according to species. In other words, any resolution of detail can be defined without the need to add additional dimensions or extra categories. The semantic space simply becomes defined by the various sub-regions. If an image of a big cat that has not been previously scaled (such as a cougar) needs to be included in the database, the classification algorithm will likely be able to identify it as a big cat even if it does not know how to distinguish a cougar. This will place the cougar's *humanness* coordinate somewhere within the big cat range and the image will still be successfully retrieved if looking for big cats, wildlife, nature, or any other category associated with that general region of the space.

Designing experiments for testing image quality can sometimes be difficult due to the nature of image-dependencies. Using the 10 perceptual semantic dimensions as a guide in image selection can help to reduce wasted research efforts. By identifying images to include in image quality experiments according to the perceptual semantic dimensions rather than intuition, the selected images will vary in the ways that are psychologically most important. This helps remove the guesswork from experimental design.

9.3 PERCEPTUAL DIMENSION DEFINITIONS

Ten perceptual semantic dimensions have been named that can account for all 34 fundamental categories, however these dimensions have not been formally defined. Following are definitions for each of the 10 perceptual dimensions.

Humanness	The degree to which living creatures in an image, if any, exhibit human characteristics.
Artificialness	The degree to which objects in an image are manmade.
Perceived Proximity	The distance from the observer's perceived position in the scene to the focal point of the image.
Candidness	The degree to which subjects in the image are aware of the picture taking process. When the *humanness* dimension is zero (no living creatures), then *candidness* is undefined. This is similar to the way *hue* does not have meaning when *chroma* is zero.
Wetness	The degree to which the presence of water in an image is considered a significant image element.
Architecture	The degree to which architecture or architectural elements are considered a significant image element.
Terrain	The primary type of terrain in a scene such as *floral, green foliage, dirt, sandy, rocky, and snowy*. If a scene contains more than one type of terrain, then the most significant terrain is used.
Activeness	The degree of physical activity in a scene. *Activeness* does not need to apply only to living creatures. Concepts such as weather can create activity as well so an image of a tornado, for example, would have a high level of *activeness*.
Lightness	The degree to which an image is naturally light as in *night* vs. *day*.
Relative Age	The age of a person or object relative to its typical lifespan. For people, a child would have a low *relative age* and an person in their 80's would have a high *relative age*. For objects such as architecture, a modern building would have a low *relative age*, a building from the late 19th century might have a mid-range *relative age*, and ancient ruins would have a high *relative age*.

10. SUMMARY

Two psychophysical experiments were conducted to identify the fundamental image categories for typical consumer imagery. Both experiments used a carefully selected set of 321 images. Experiment I was a Free Sorting Experiment where observers were asked to sort the 321 images into piles of similar images. Thirty observers participated in the Free Sorting Experiment. Piles from all thirty observers were compiled to form a similarity matrix. Experiment II was a Distributed Experiment conducted over the internet which used the method of triads to collect similarity and dissimilarity data from the 321 images. Due to the large number of images included in the experiment, the method of non-repeating random paths was employed to reduce the number of required responses. Approximately 9,152 observers participated in the Distributed Experiment.

Both experiments were analyzed using multidimensional scaling and hierarchical cluster analysis. The Free Sorting Experiment was also analyzed with dual scaling. The two-dimensional and three-dimensional global configurations from the multidimensional scaling were examined and semantic trends were observed in these configurations that led to the identification of the first set of candidate categories. The dimensions that resulted from dual scaling were also examined for semantic trends by inspecting the images that appeared at either extreme of each dimension. This process yielded the next set of candidate categories.

Hierarchical cluster trees were calculated through hierarchical cluster analysis. Each hierarchical cluster tree was forced into a fixed number of clusters ranging from two through twenty clusters. Each cluster was examined for semantic cohesion and if a cluster appeared to contain primarily a single semantic theme and also appeared to be stable across multiple linkage methods, then it was added to a list of candidate categories. This process continued as the number of fixed clusters increased until the results were no longer interpretable. The result of this analysis was a third set of candidate categories.

Six-cluster groupings were selected from the hierarchical cluster analysis to form local groupings in the multidimensional scaling configurations. The centroid for each of the six clusters were calculated and a fixed radius from each centroid was defined. Multidimensional scaling was recalculated for each of the six local regions using all images contained within each of the six circular boundaries. This process yielded six new two-dimensional configurations for each experiment. These local configurations were examined for additional semantic trends that were previously obscured in the global configuration and a fourth set of candidate categories was formed.

The four sets of candidate categories were compiled and a set of 34 categories that proved to be stable across multiple methods of analysis was formed. These categories are the proposed final set of fundamental semantic categories for typical consumer imagery.

10.1 CONCLUSIONS

A multidimensional perceptual image semantic space has been suggested and advantages to utilizing such a structure have been outlined. The 34 fundamental categories that have been identified can be represented by 10 perceptual dimensions which describe the underlying perceptions that lead to categorical assignments. The 10 perceptual dimensions are *humanness, artificialness, perceived proximity, candidness, wetness, architecture, terrain, activeness, lightness, and relative age.*

The proposed semantic space could enable imaging scientists to solve some of the imaging problems described in the introduction. It could simplify the problem of image-dependent image processing algorithms by providing a simple way to communicate image semantics without the need for making finite categorical assignments. Coordinates in the perceptual semantic space are all that are needed to identify the multitude of categories to which an image may belong. A perceptual semantic space also has the potential to greatly simplify the task of image retrieval since any one point in the multidimensional semantic space can represent several categories simultaneously. Depending on the search parameters, it would only be necessary to find all images contained within a sub-region of the space. To search for a new category not originally defined for the database, it would only be necessary to determine which sub-region would best represent that category. Finally, designing experiments in which the results can be shown to be image dependent can be streamlined by utilizing the 10 perceptual semantic dimensions.

10.2 FUTURE WORK

A perceptual image semantic space has been suggested and a set of 10 fundamental dimensions has been identified. However, before such a space could be utilized, it will be necessary to conduct additional experiments to scale each of the dimensions in order to quantify them so that an image's coordinates can be calculated.

This will require a separate psychophysical experiment for each dimension that needs to be scaled. It will also be necessary to find image descriptors which can be correlated with each of the dimensions and calculated automatically.

An exciting aspect of the proposed semantic space is that it could potentially evolve into a universal image semantic description that is not application dependent. For example, consumer imagery and photojournalism are distinctly different types of photography. However, all 10 of the perceptual dimensions could also apply to photojournalism. Perhaps additional dimensions could be uncovered that will distinguish between these types of images. In other words, perhaps there are *image type* perceptions that could provide one or more new dimensions that will identify what type of image it is. There is no doubt that determination of an *image type* will often be context dependent. Perhaps *context* itself is another perceptual dimension.

It was stated earlier that image semantics are application dependent. The universal framework would not only enable encoding of multiple, hierarchical semantic categories through the use of perceptual dimensions, but would also enable definition of multiple semantic meanings for a single image depending on the context (the *image type* perceptions) while still only requiring a small number of coordinates to be stored. Ultimately, through cooperative effort between researchers in multiple disciplines, a more complete semantic space could be constructed.

APPENDIX A

Final 321 Images Selected for Psychophysical Experiments

FIGURE A-1: Final 321 images selected for use in the psychophysical experiments.

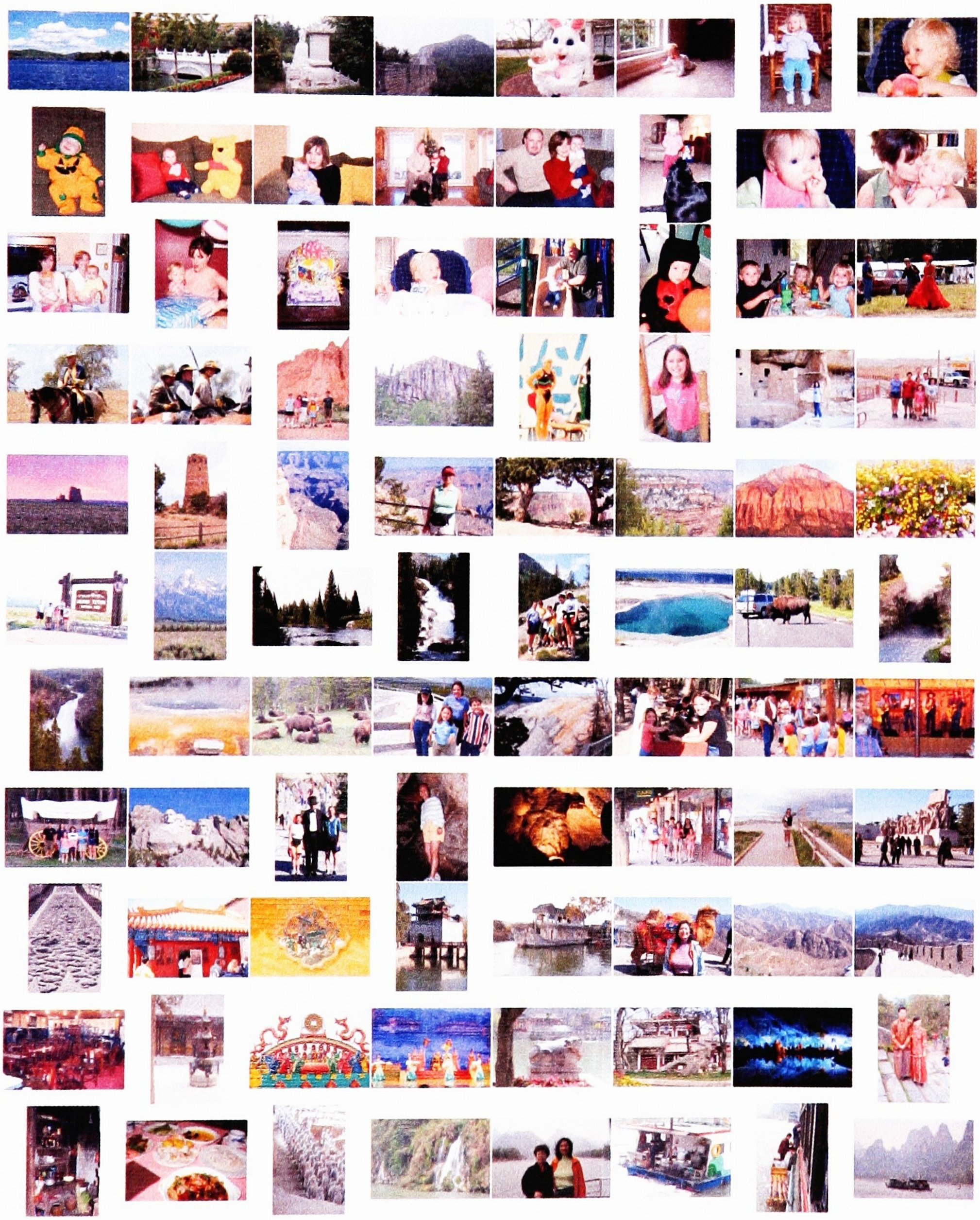


FIGURE A-1 (cont.): Final 321 images selected for use in the psychophysical experiments.

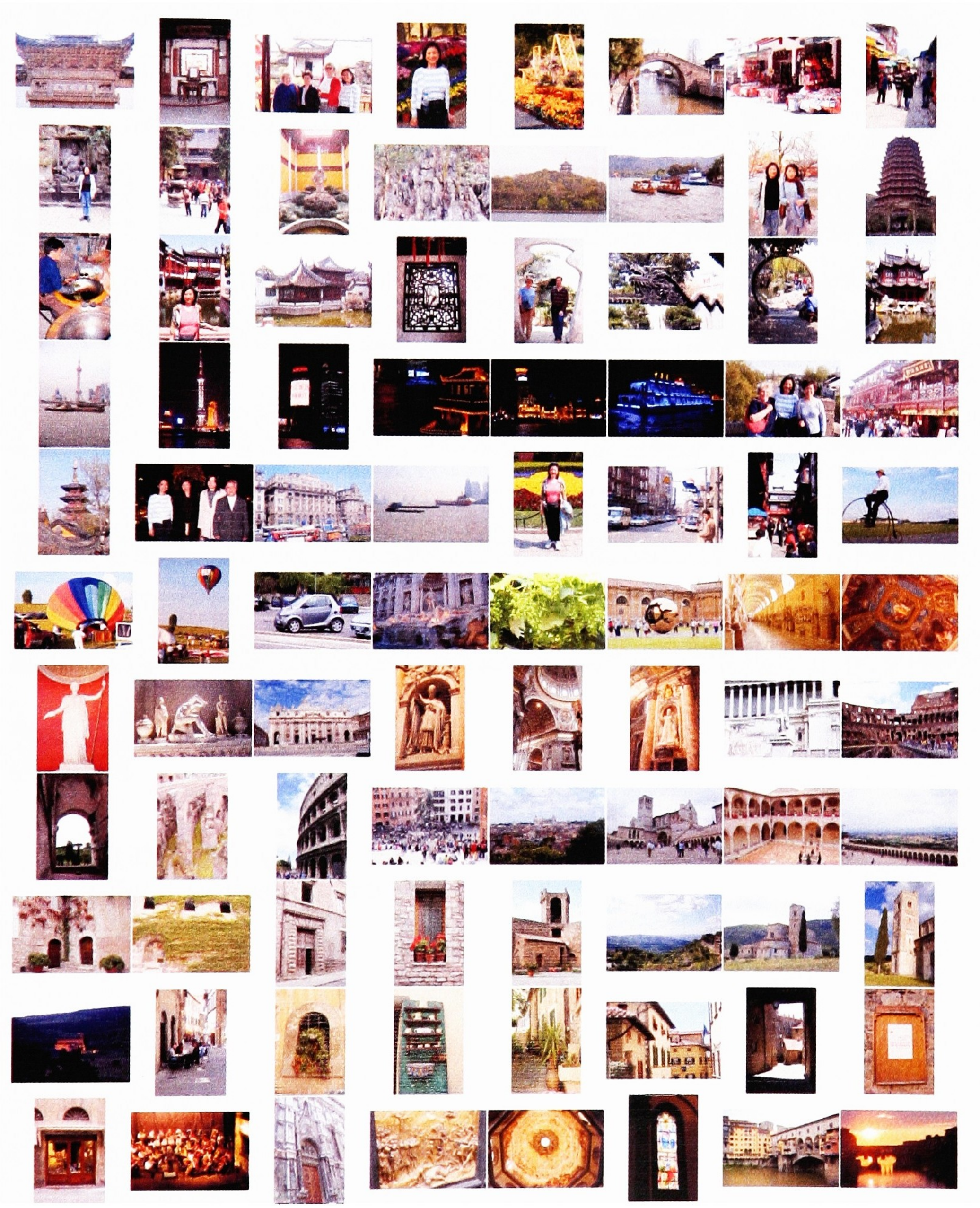


FIGURE A-1 (cont.): Final 321 images selected for use in the psychophysical experiments.

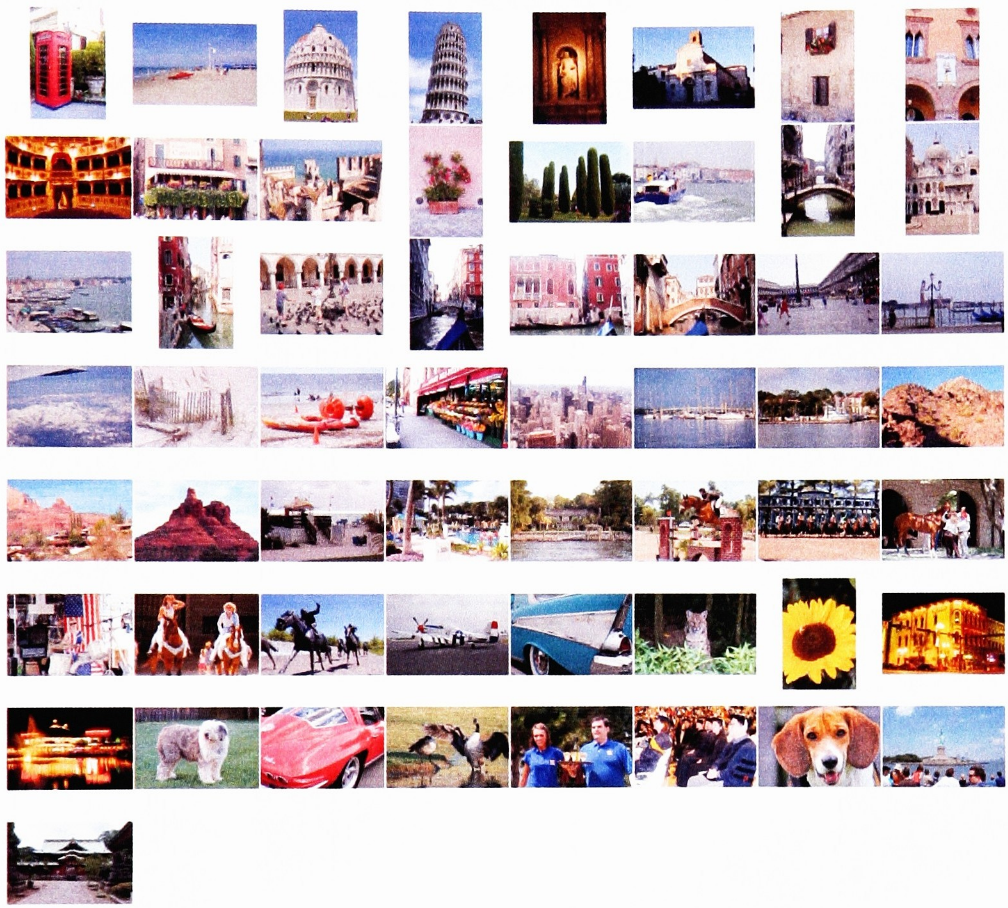


FIGURE A-1 (cont.): Final 321 images selected for use in the psychophysical experiments.

APPENDIX B

Sheppard Diagrams for the Distributed Experiment

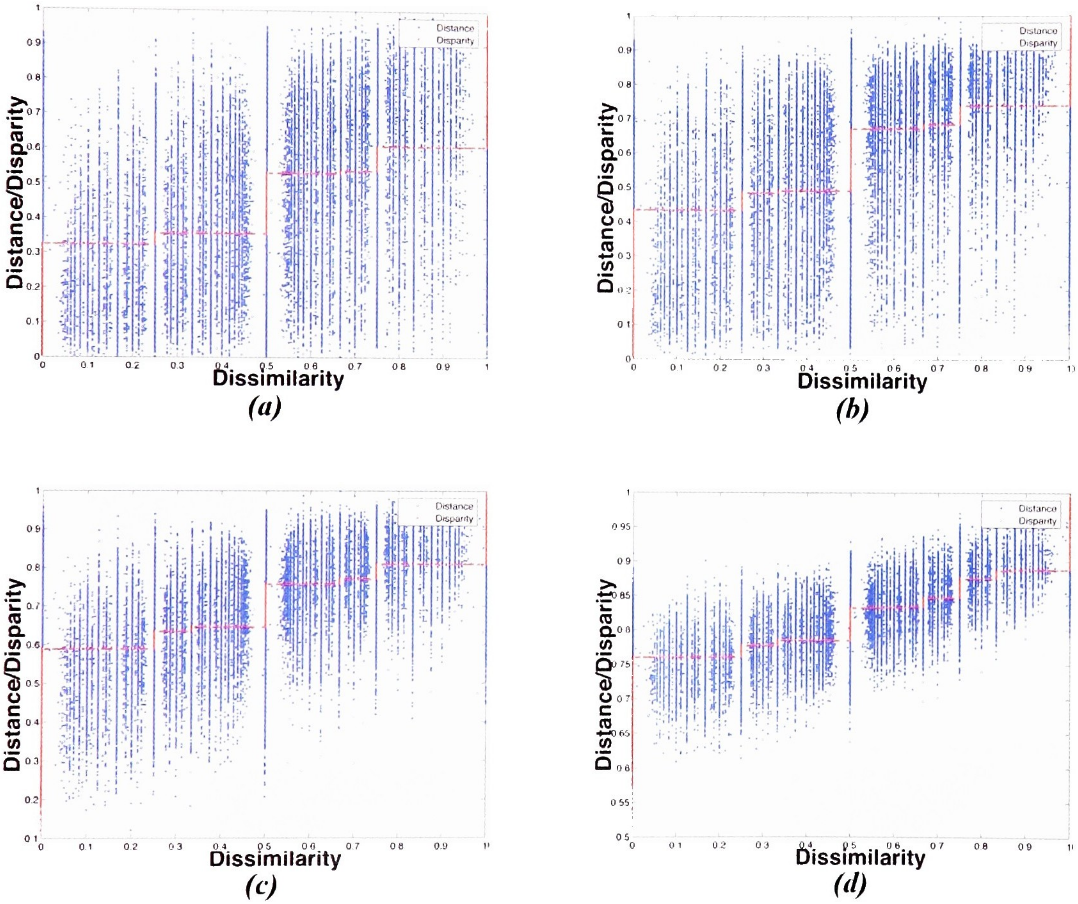


FIGURE B-1: Sheppard diagrams for the (a) 2-dimensional, (b) 3-dimensional, (c) 9-dimensional, and (d) 52-dimensional multidimensional scaling configurations.

APPENDIX C

Silhouette Plots

Silhouette plots were computed for 2-15 clusters for each of the linkage methods although only clusters 2-5 are reported. All silhouette plots for groupings greater than 5 clusters continued to degenerate gradually, as can be seen by the difference between the two cluster groupings and the 5 cluster groupings for each linkage method.

Free Sorting Experiment

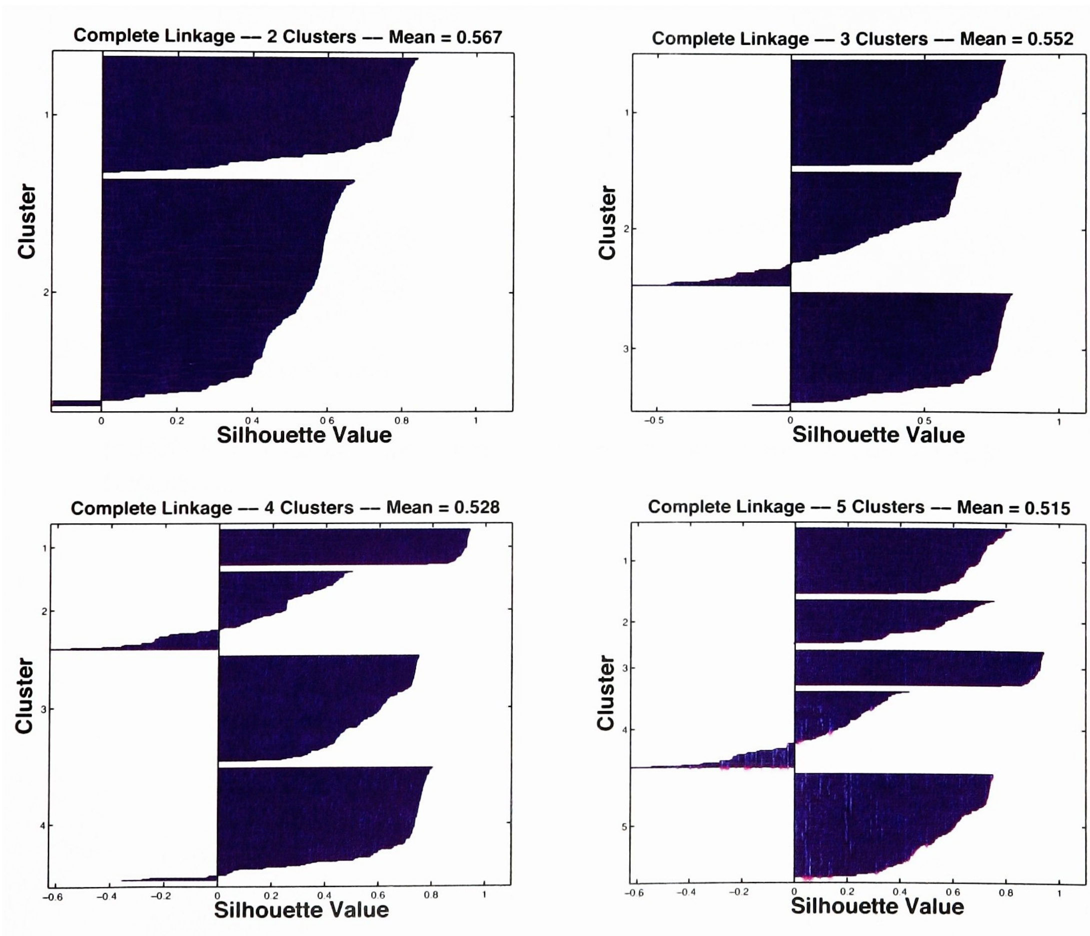


FIGURE C-1: Silhouette plots for 2-5 clusters for the Free Sorting Experiment.

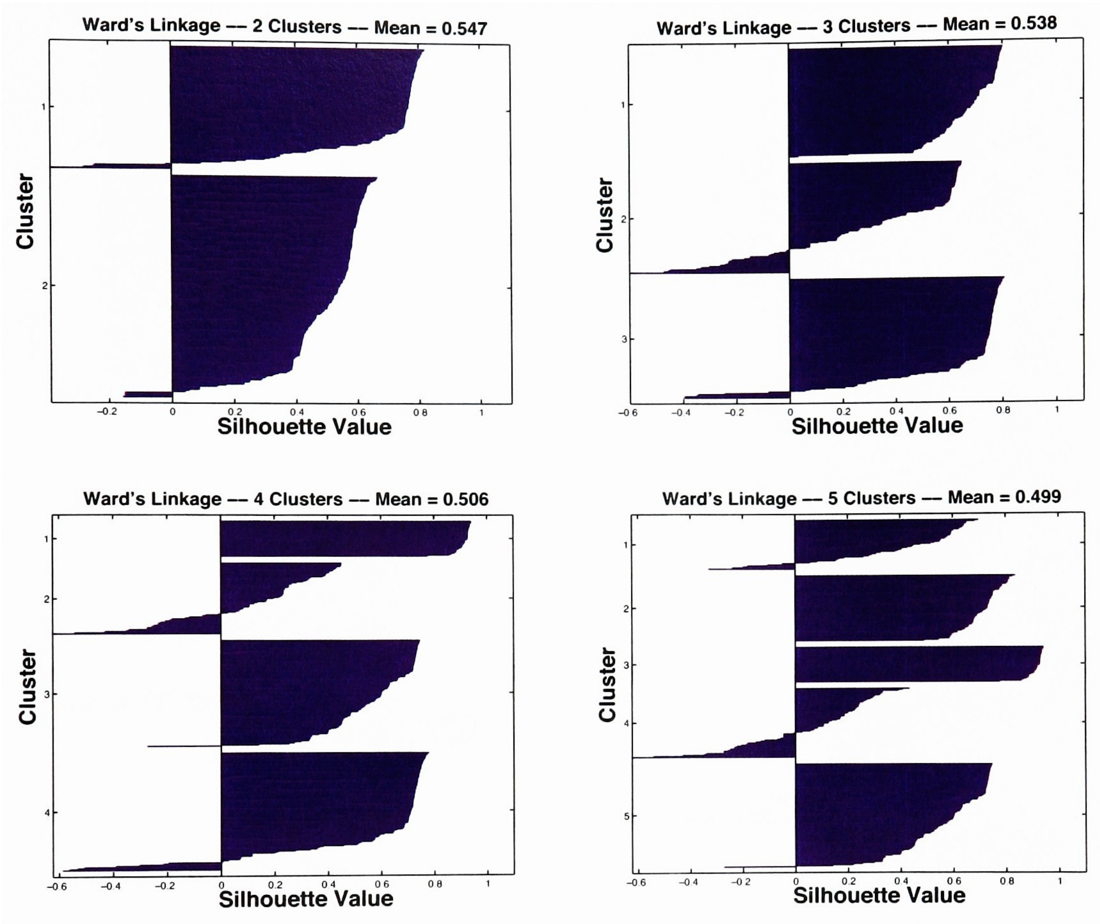


FIGURE C-1 (cont.): Silhouette plots for 2-5 clusters for the Free Sorting Experiment.

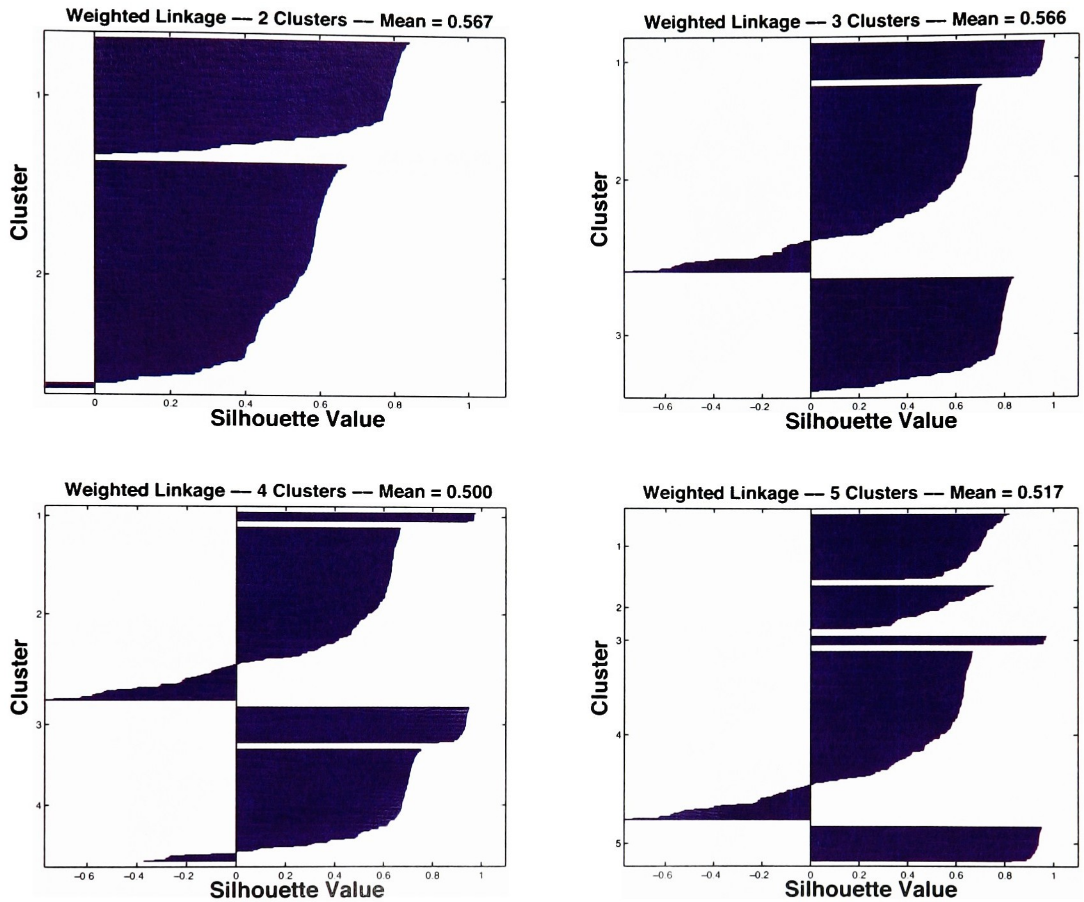


FIGURE C-1 (cont.): Silhouette plots for 2-5 clusters for the Free Sorting Experiment.

Distributed Experiment

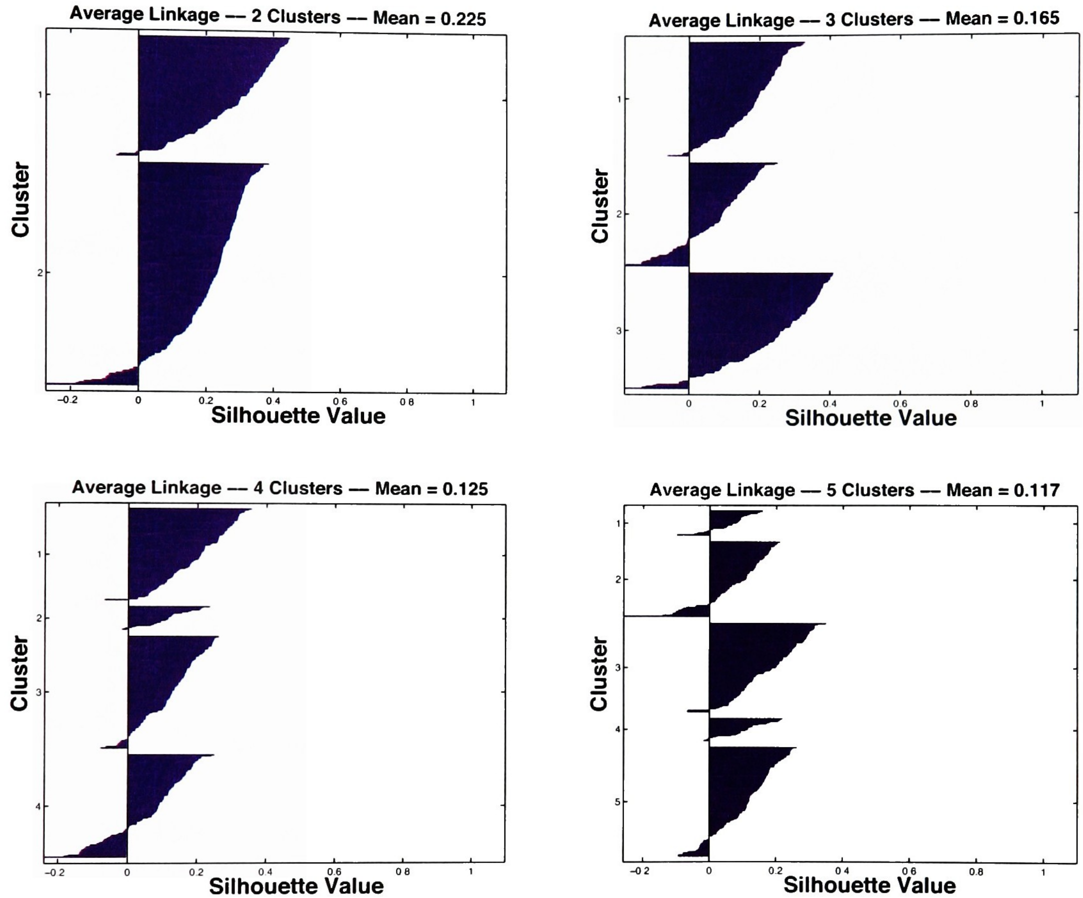


FIGURE C-2: Silhouette plots for 2-5 clusters for the Distributed Experiment.

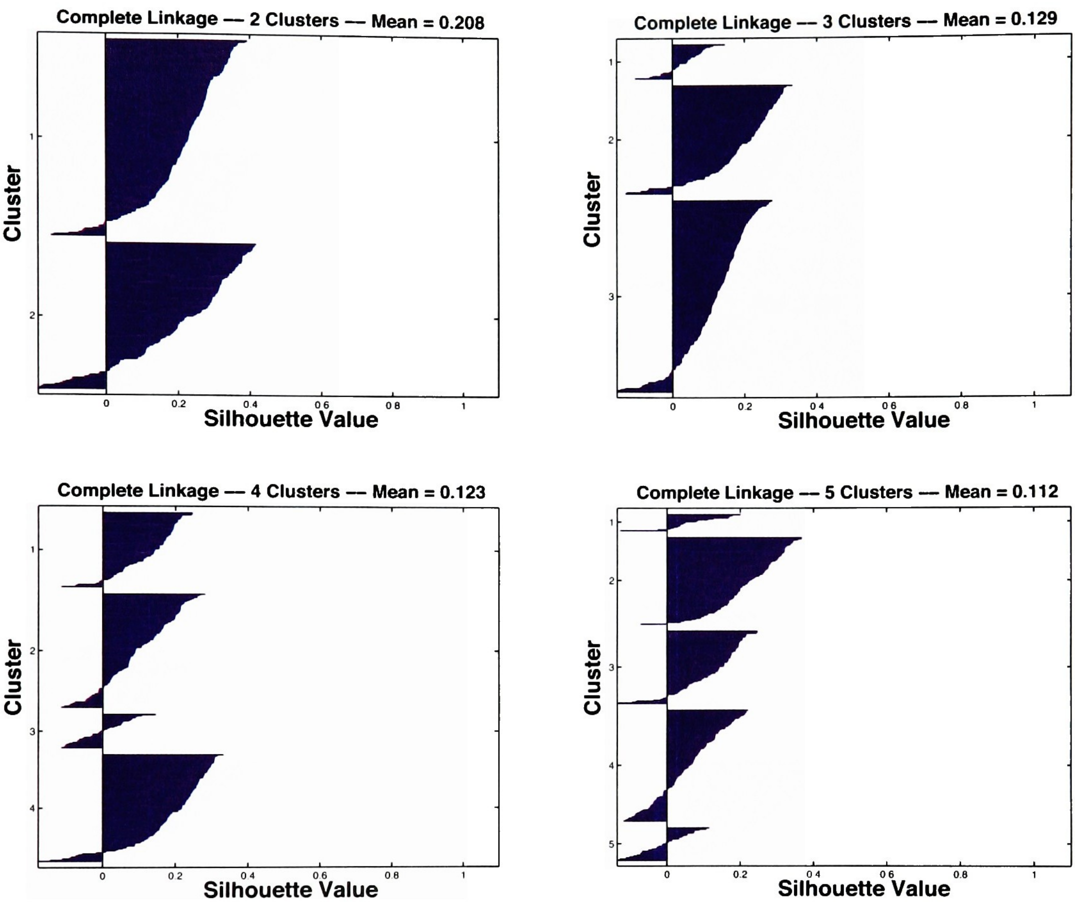


FIGURE C-2 (cont.): Silhouette plots for 2-5 clusters for the Distributed Experiment.

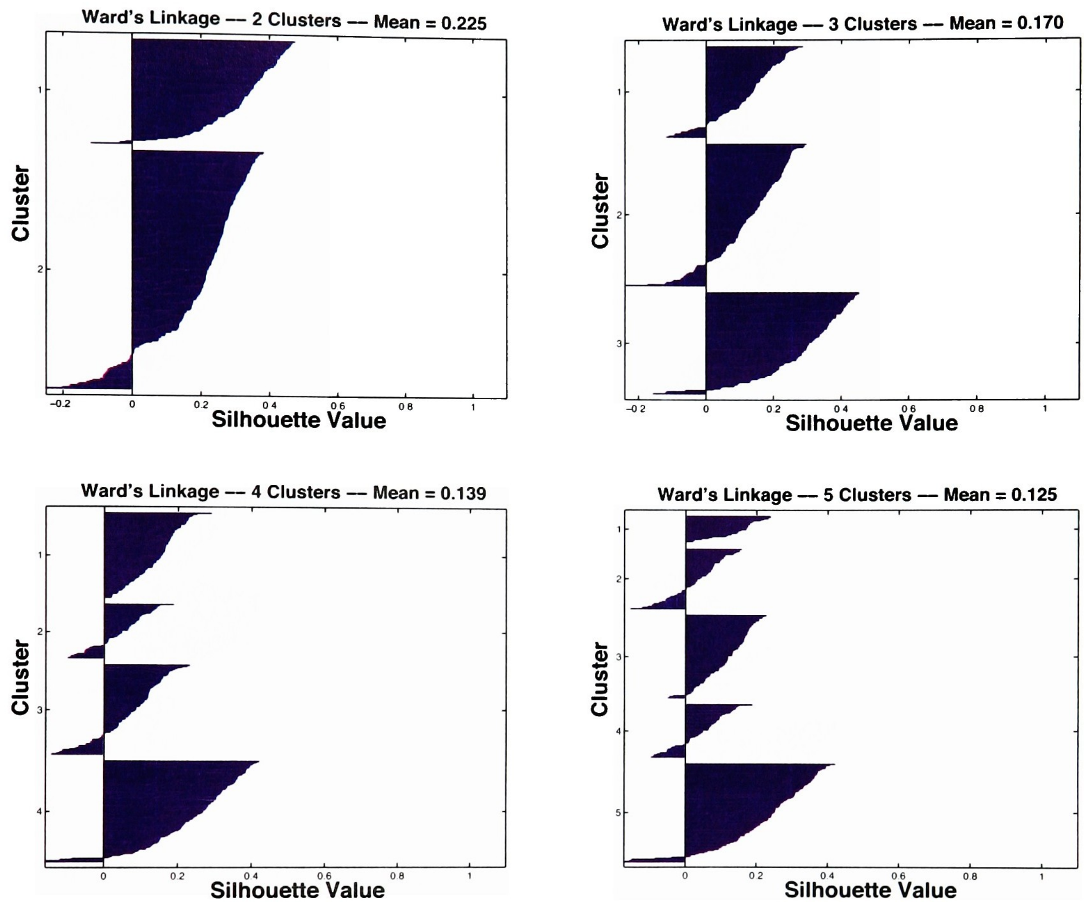


FIGURE C-2 (cont.): Silhouette plots for 2-5 clusters for the Distributed Experiment.

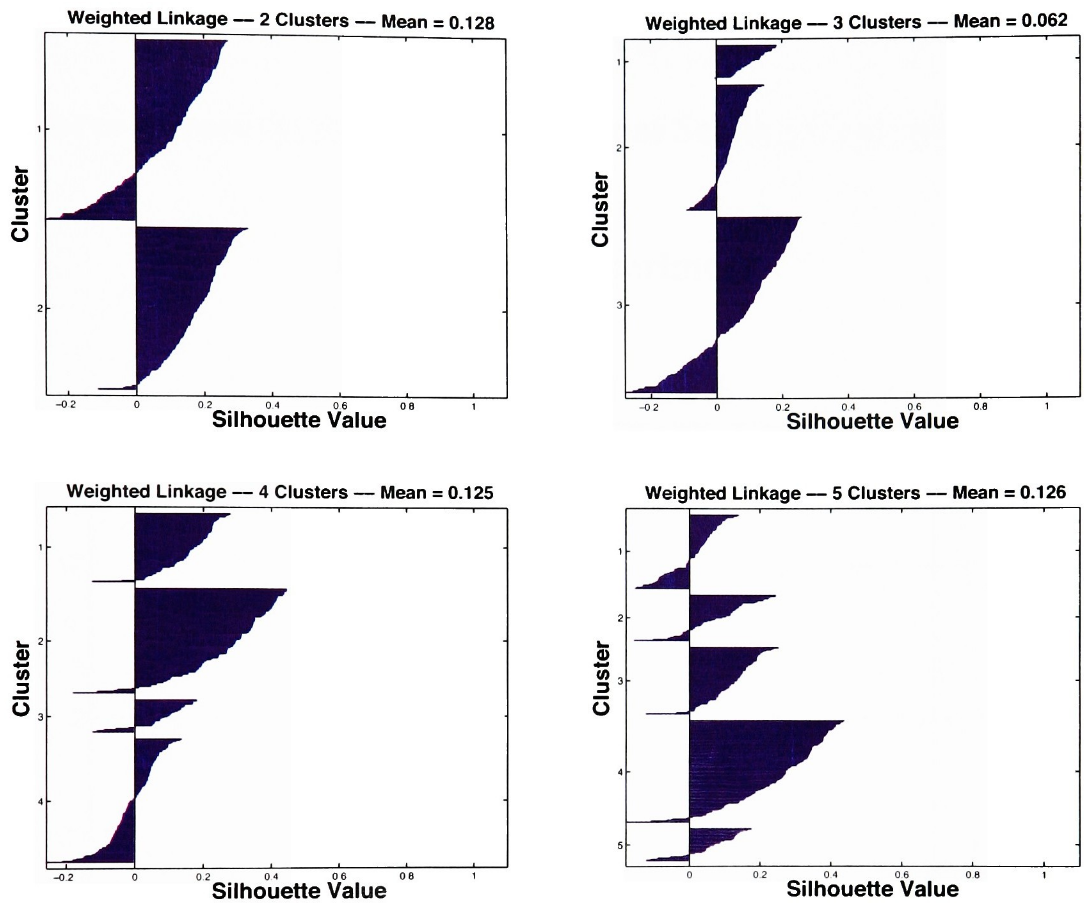


FIGURE C-2 (cont.): Silhouette plots for 2-5 clusters for the Distributed Experiment.

APPENDIX D

Three-Dimensional Multidimensional Scaling Configurations

Free Sorting Experiment

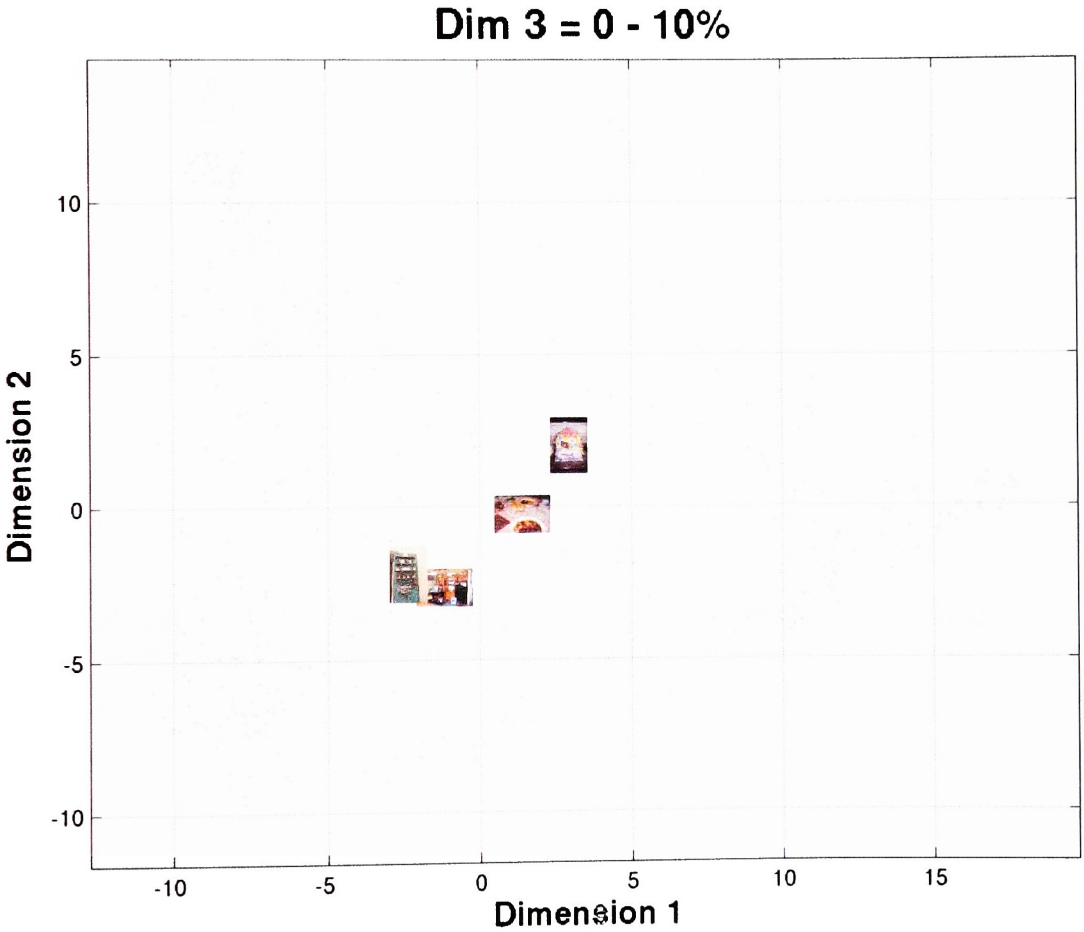


FIGURE D-1: Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

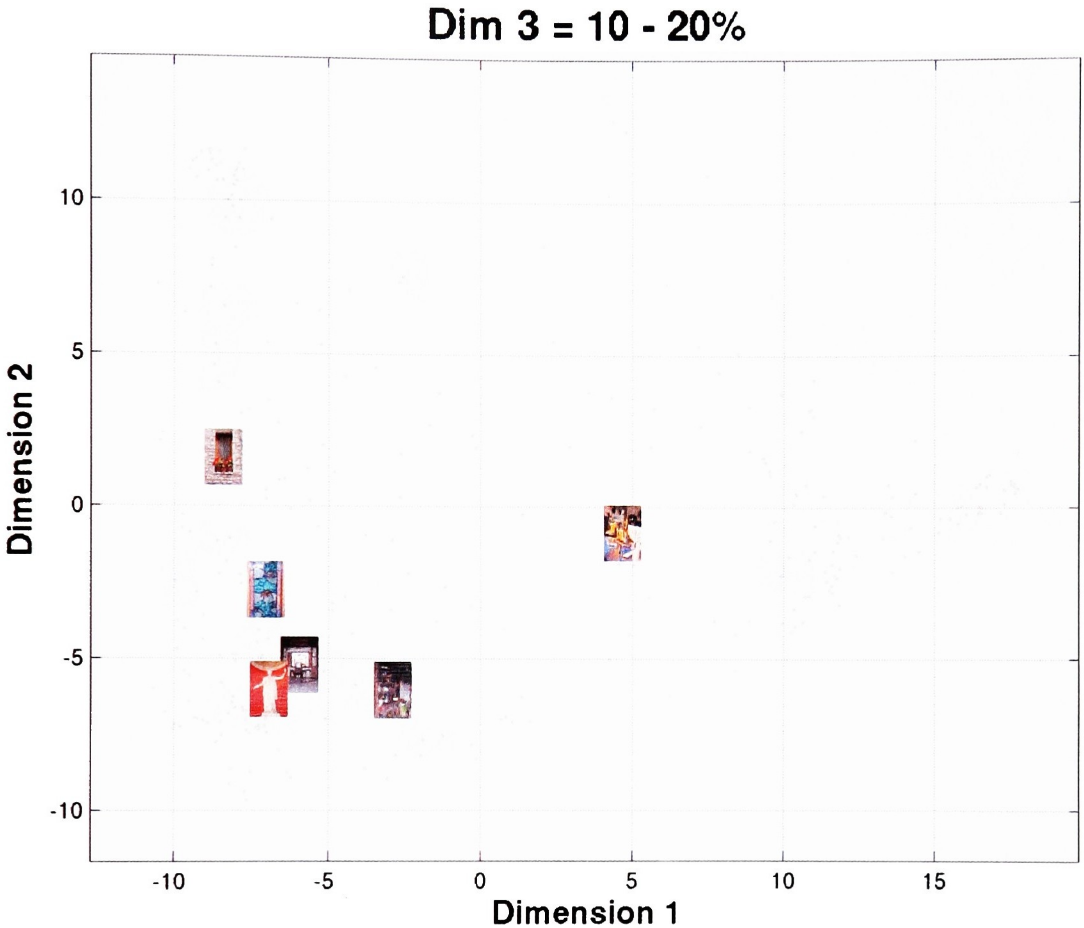


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

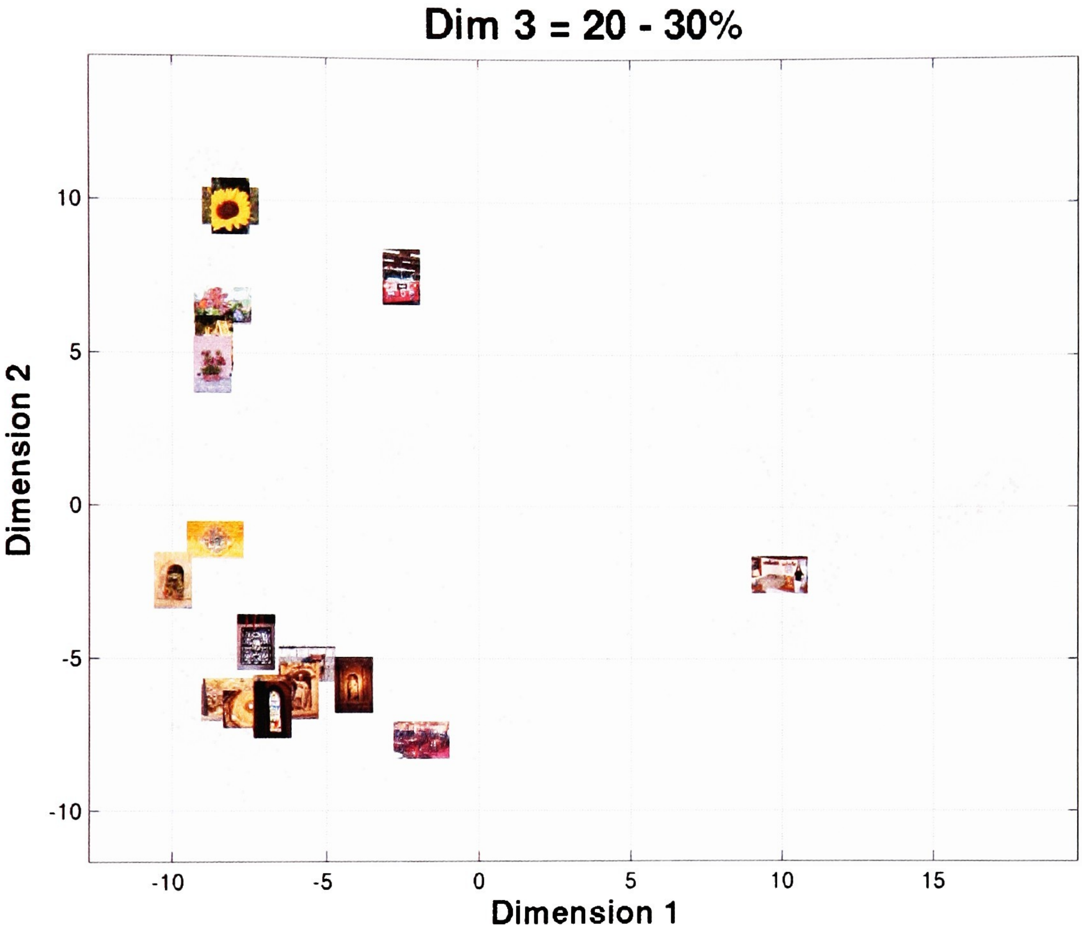


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

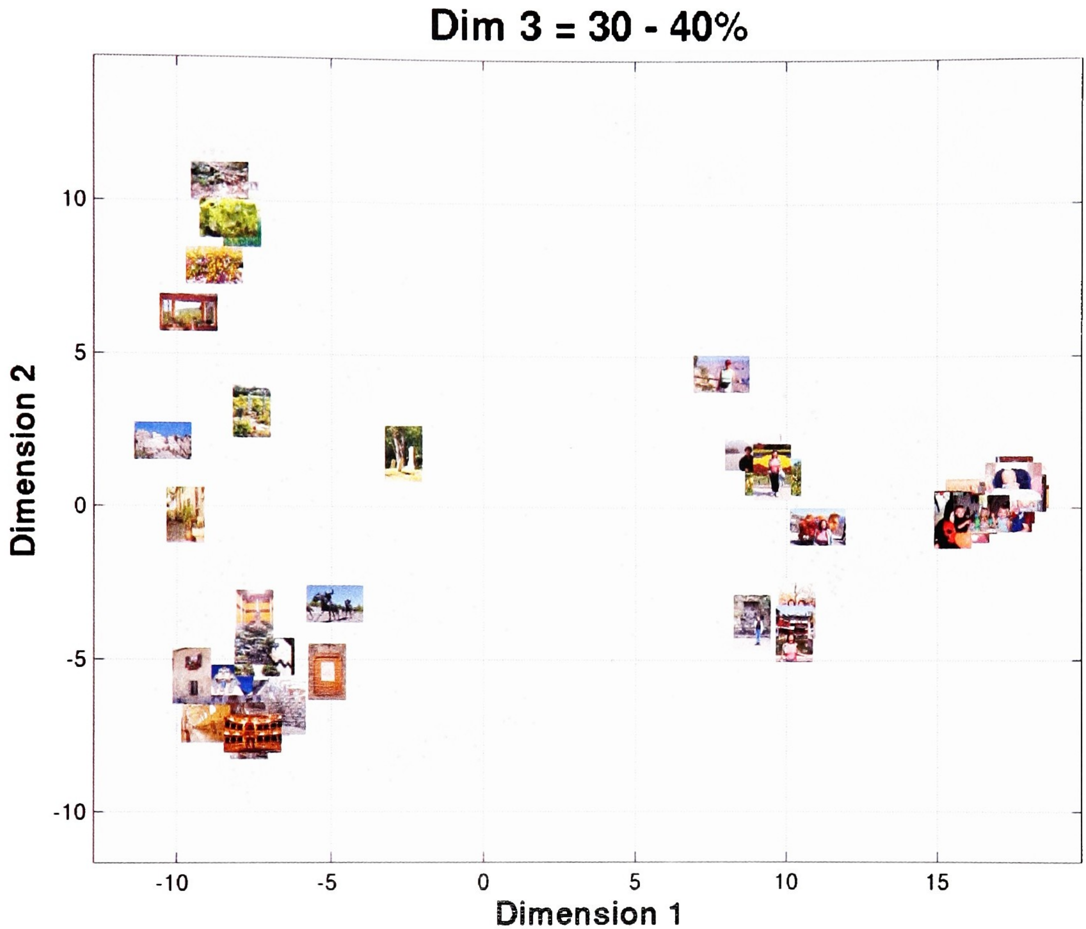


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

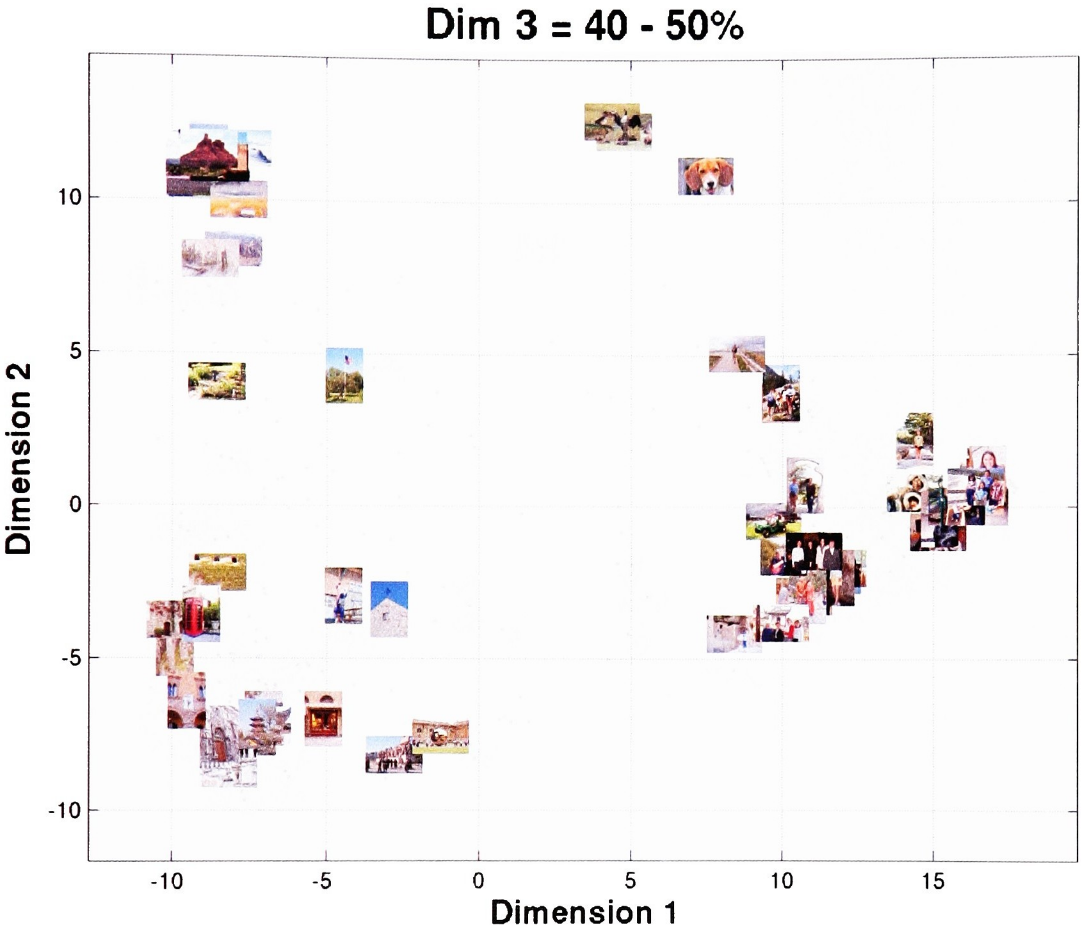


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

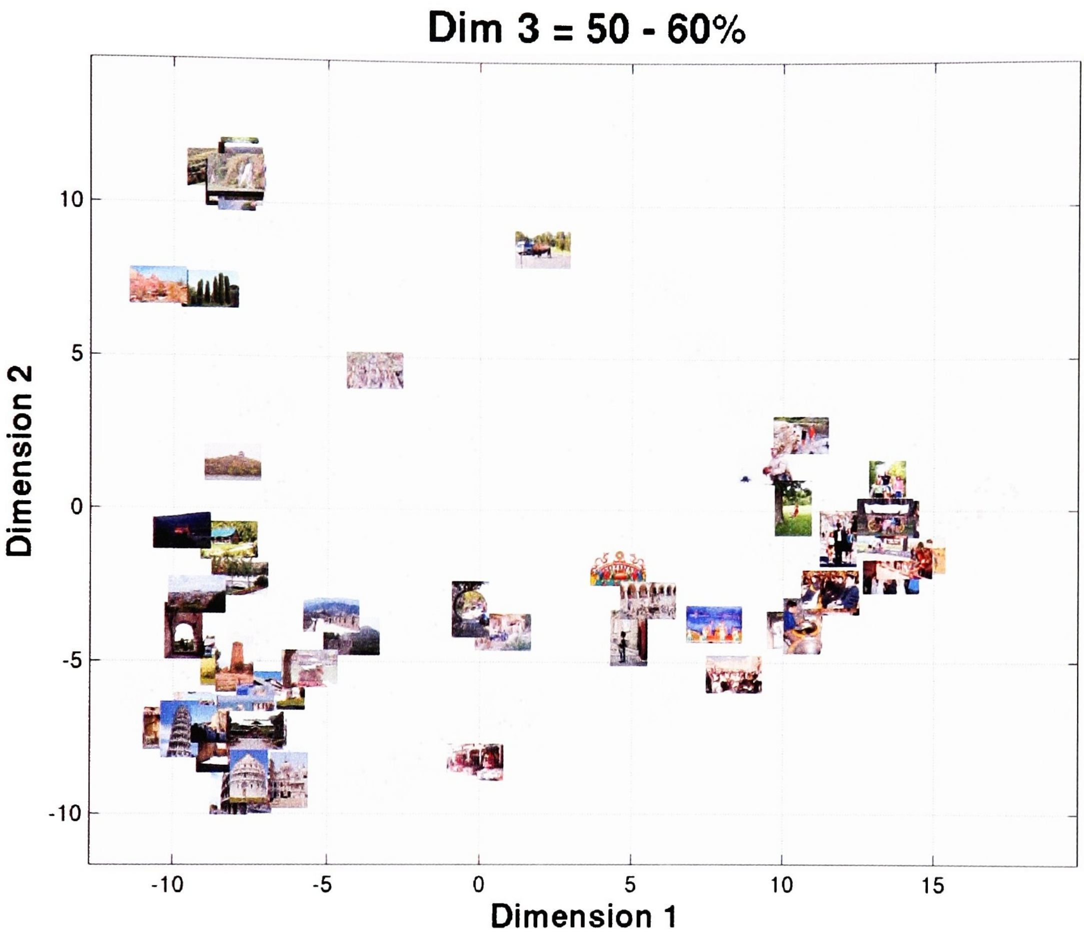


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

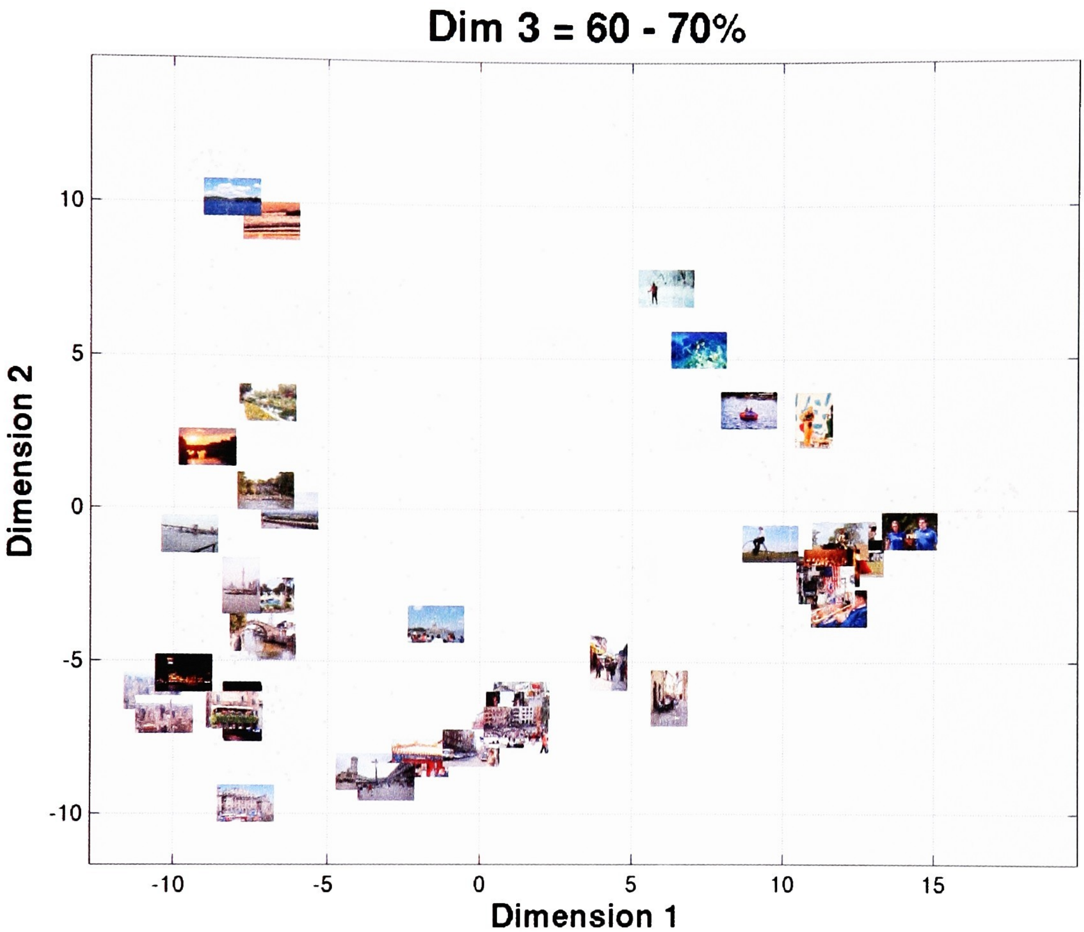


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

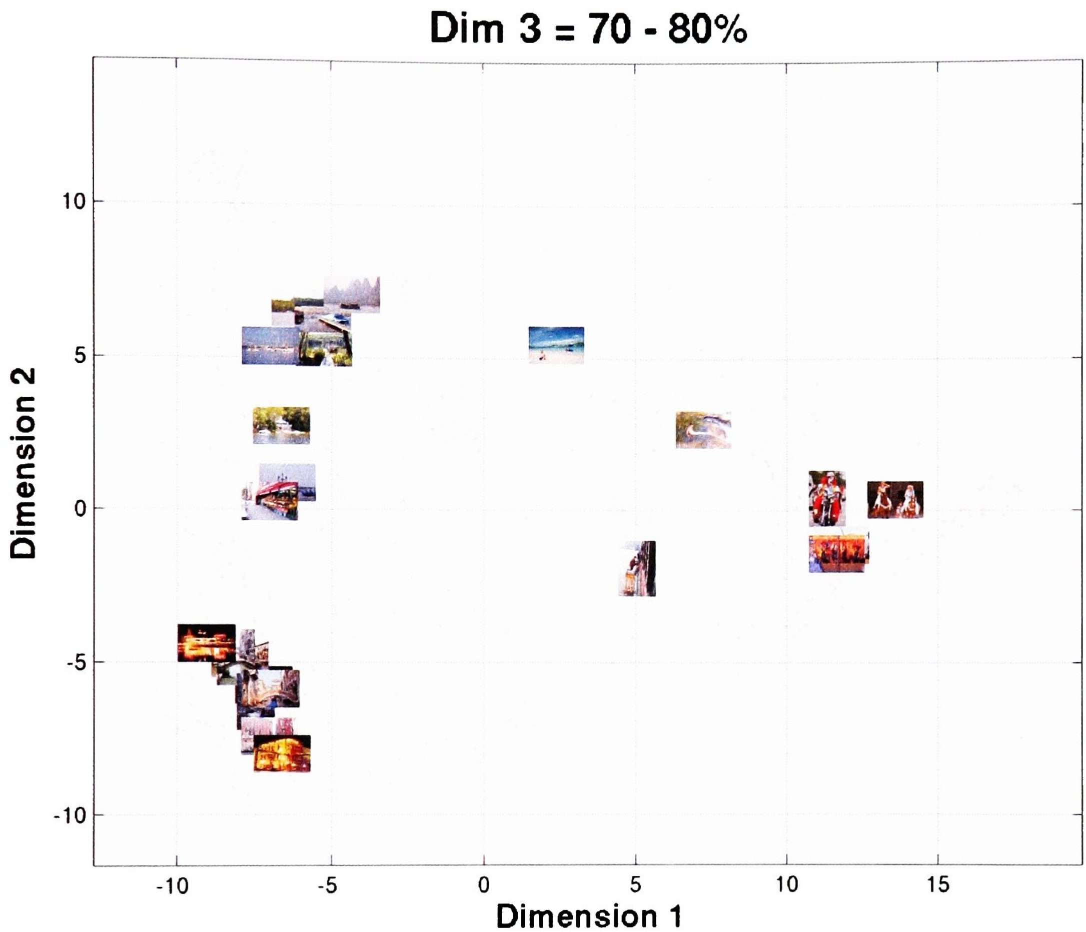


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

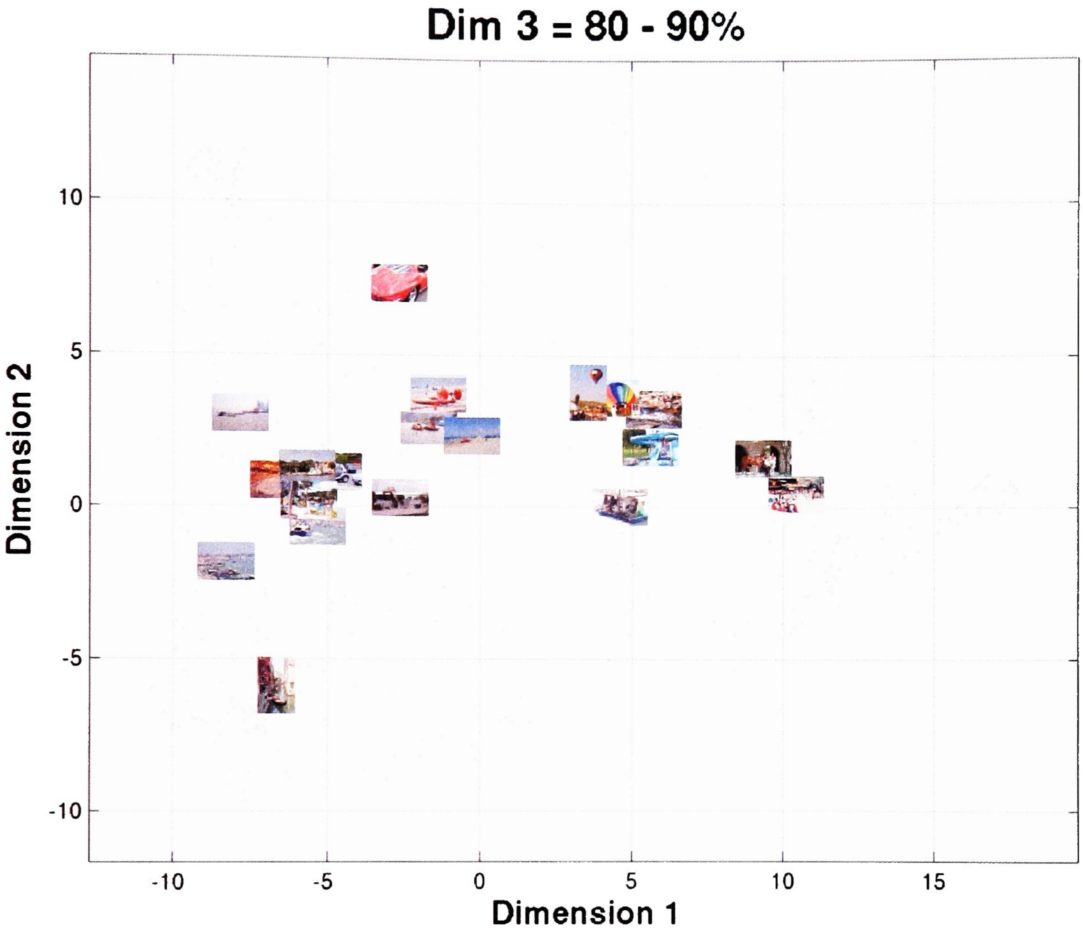


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

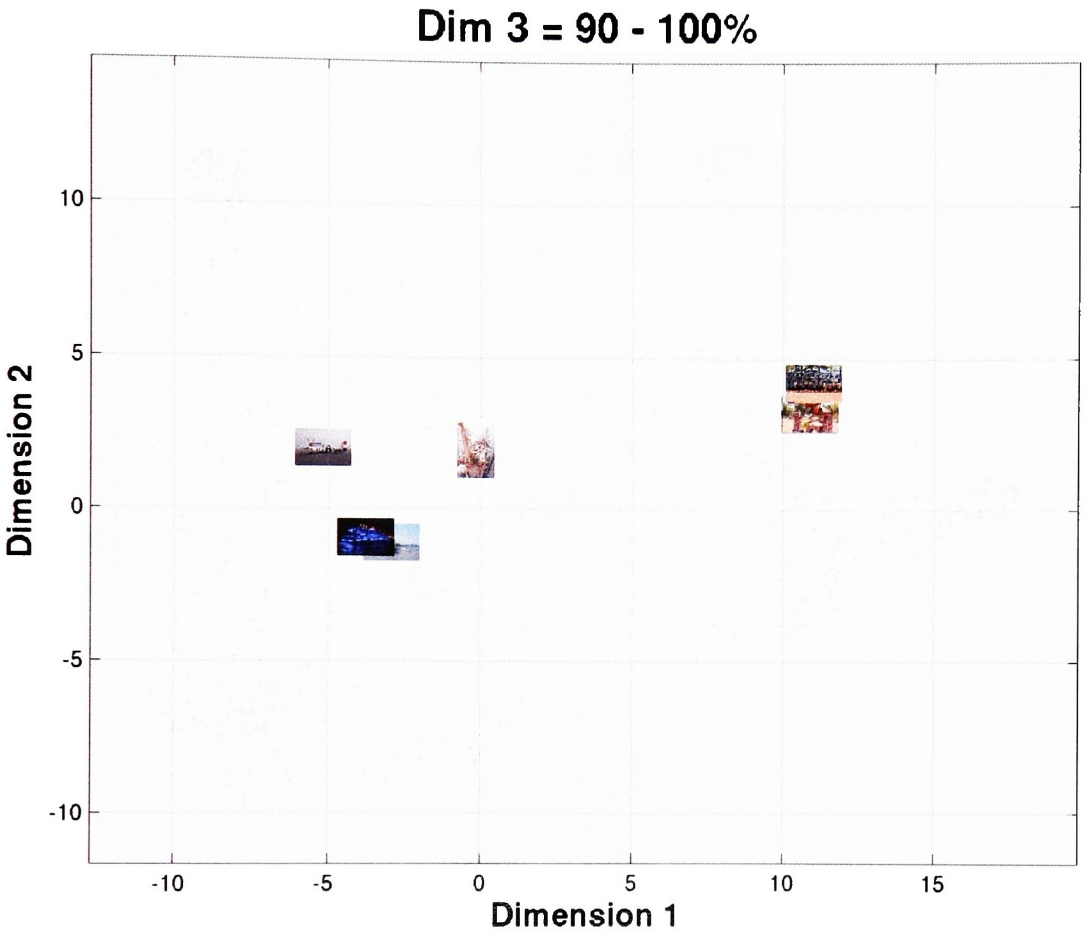


FIGURE D-1 (cont.): Three-dimensional multidimensional scaling configuration for the Free Sorting Experiment.

Distributed Experiment

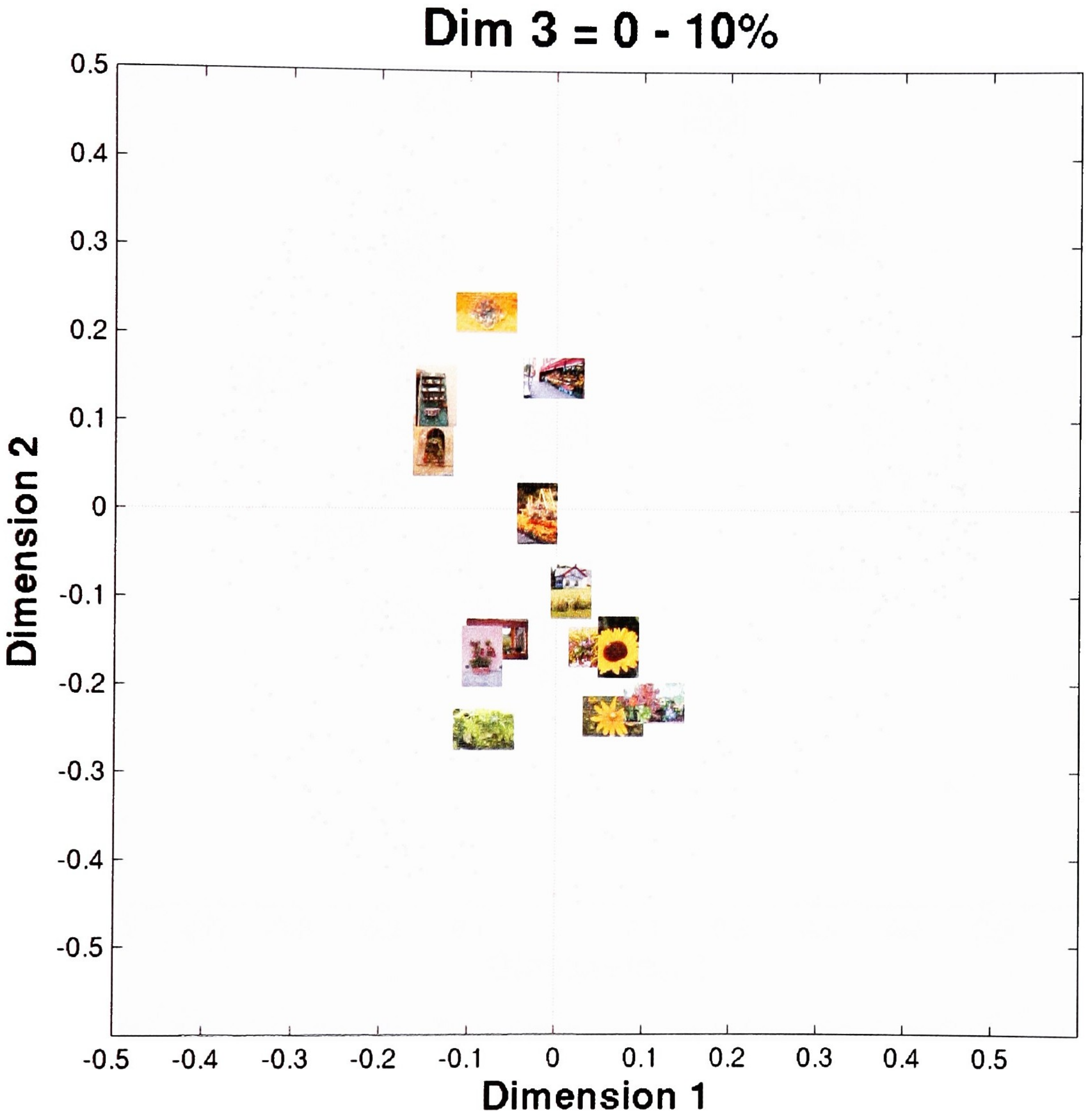


FIGURE D-2: Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

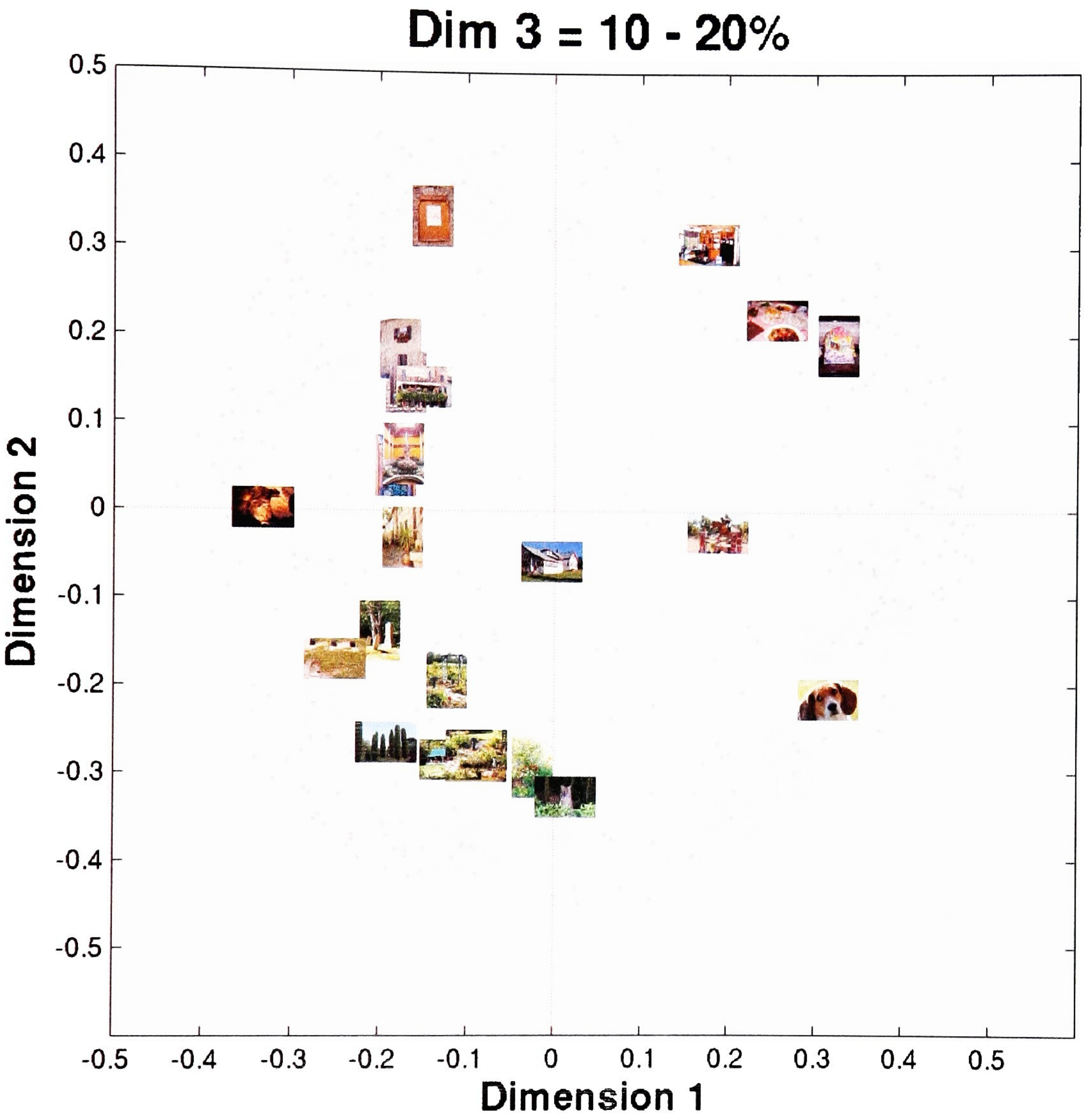


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

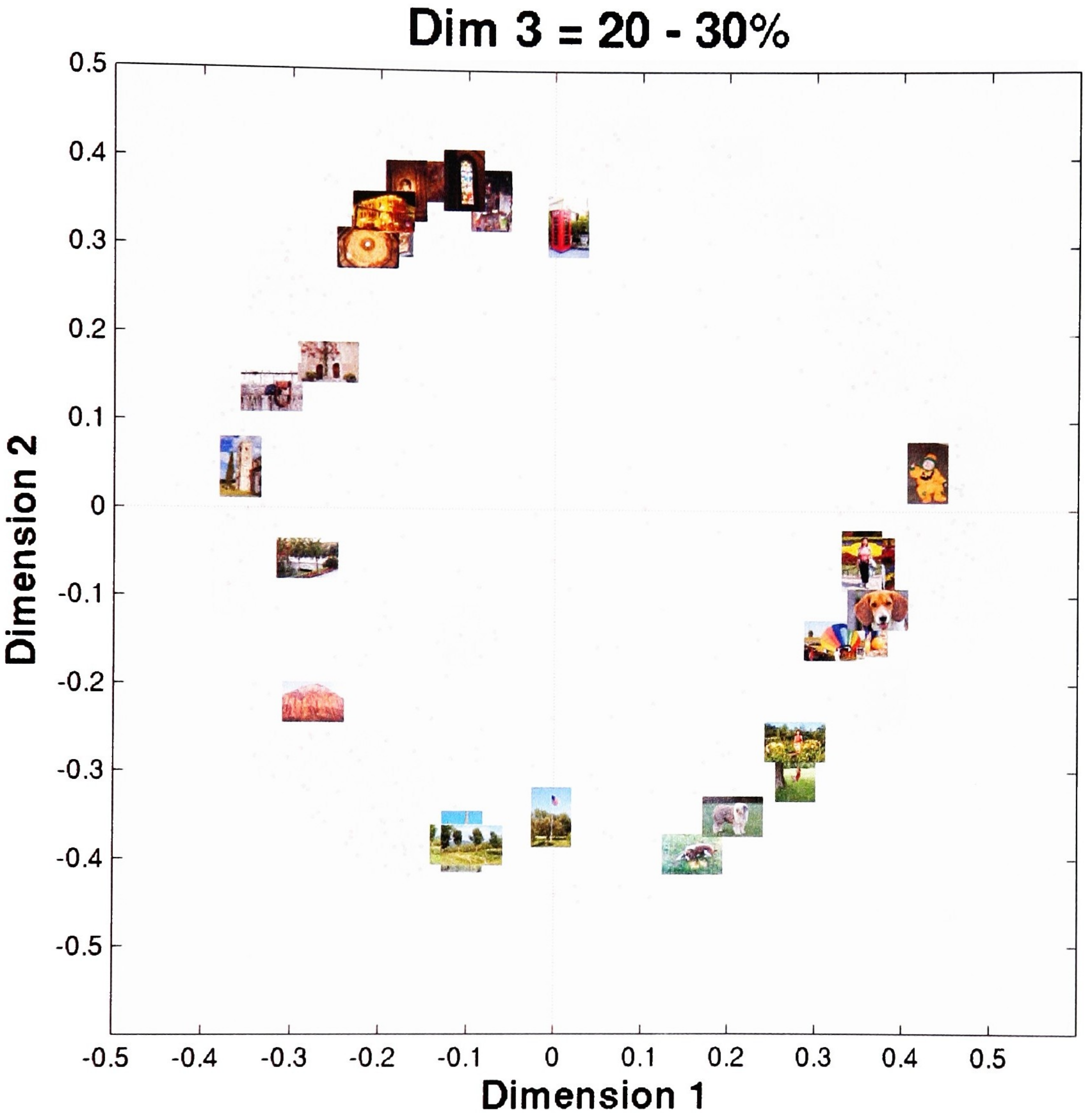


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

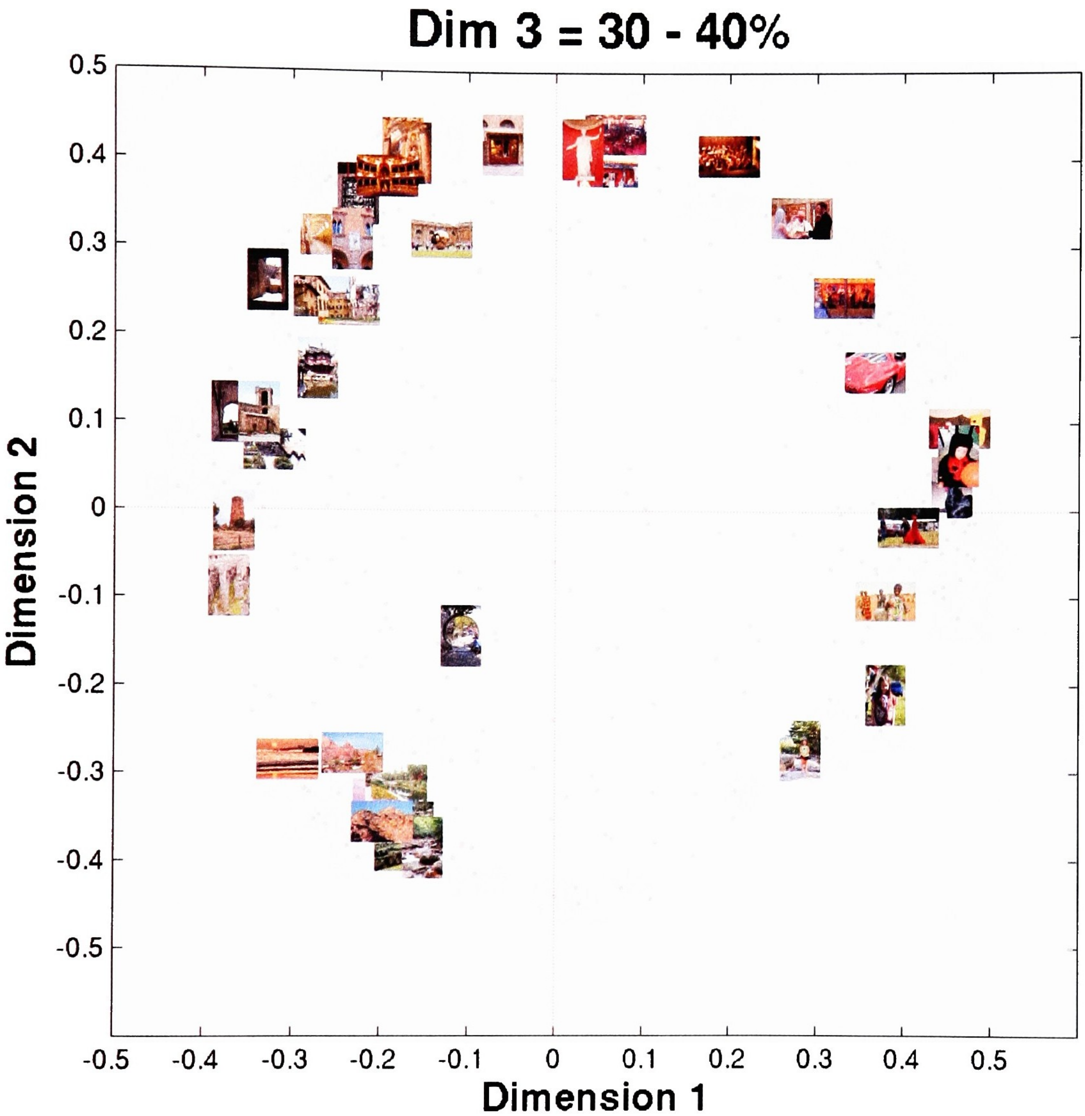


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

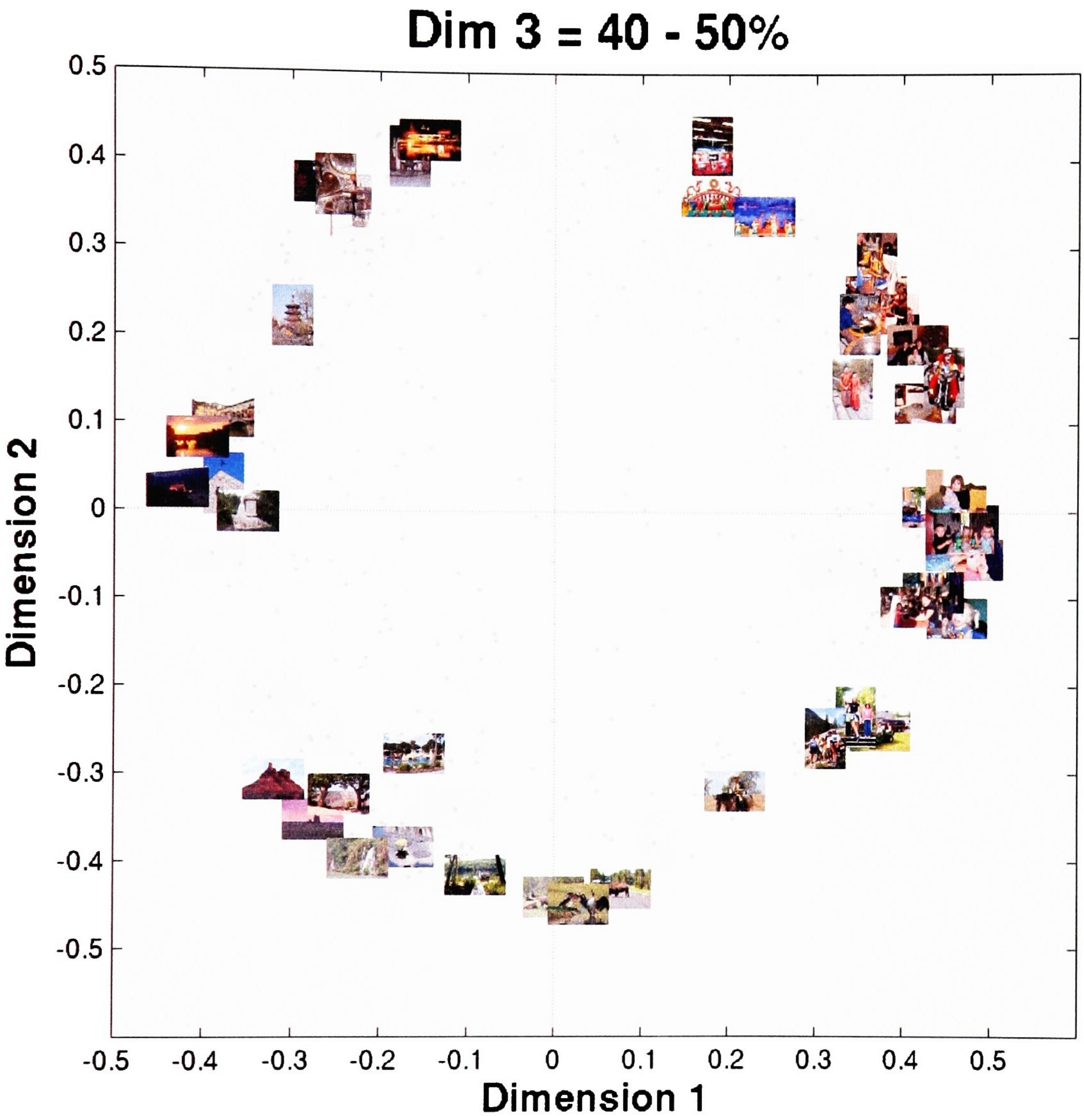


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

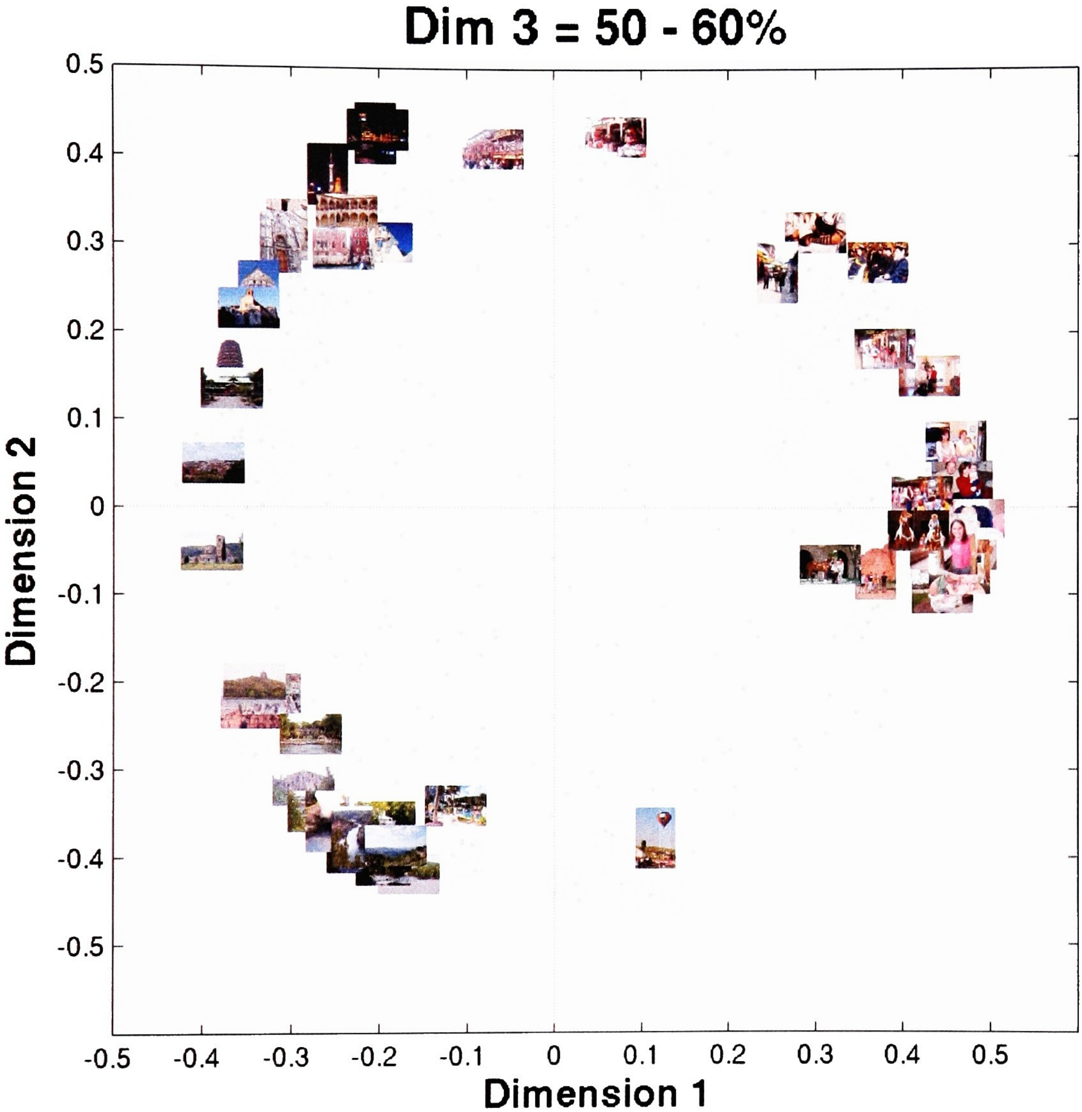


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

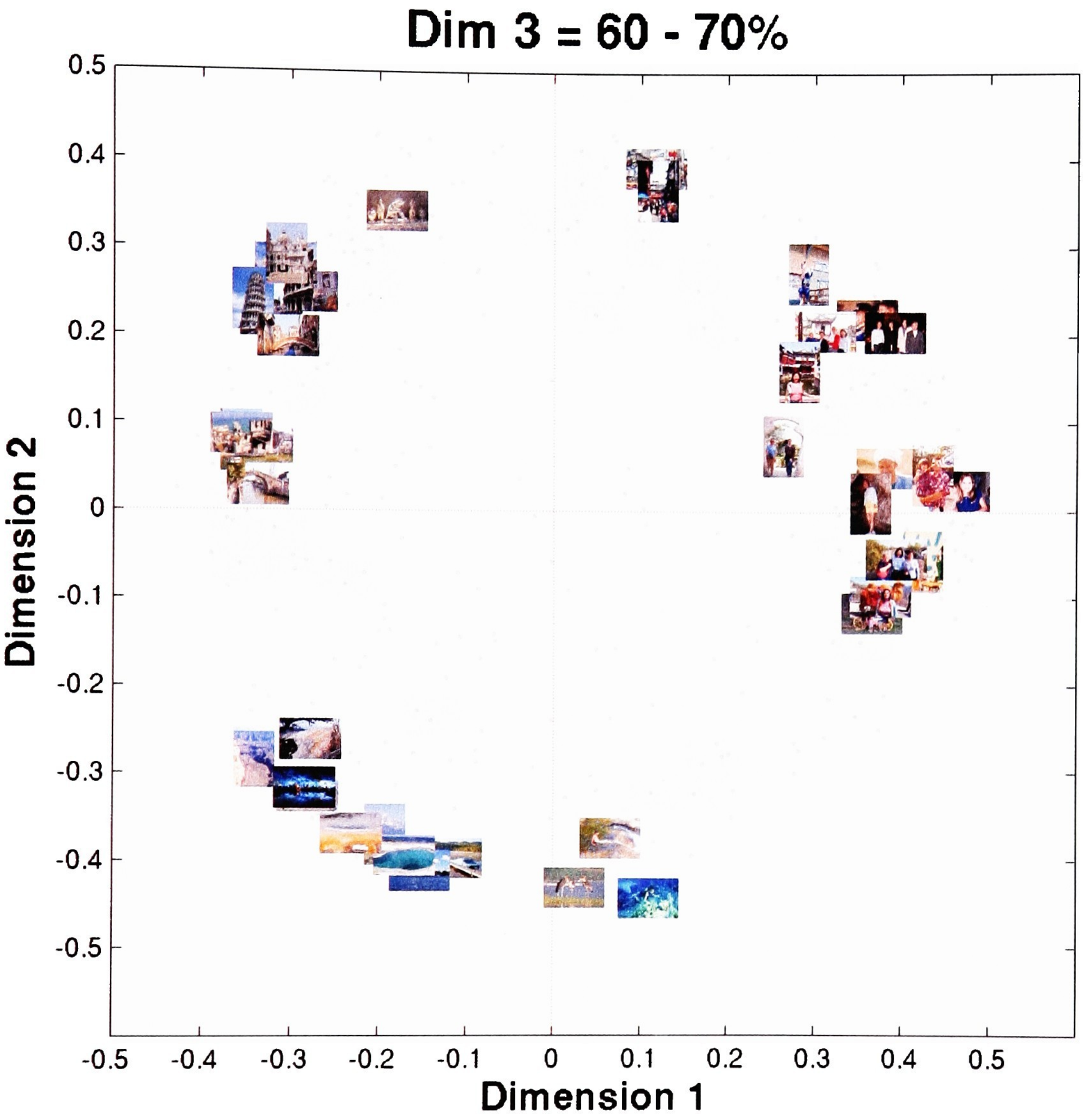


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

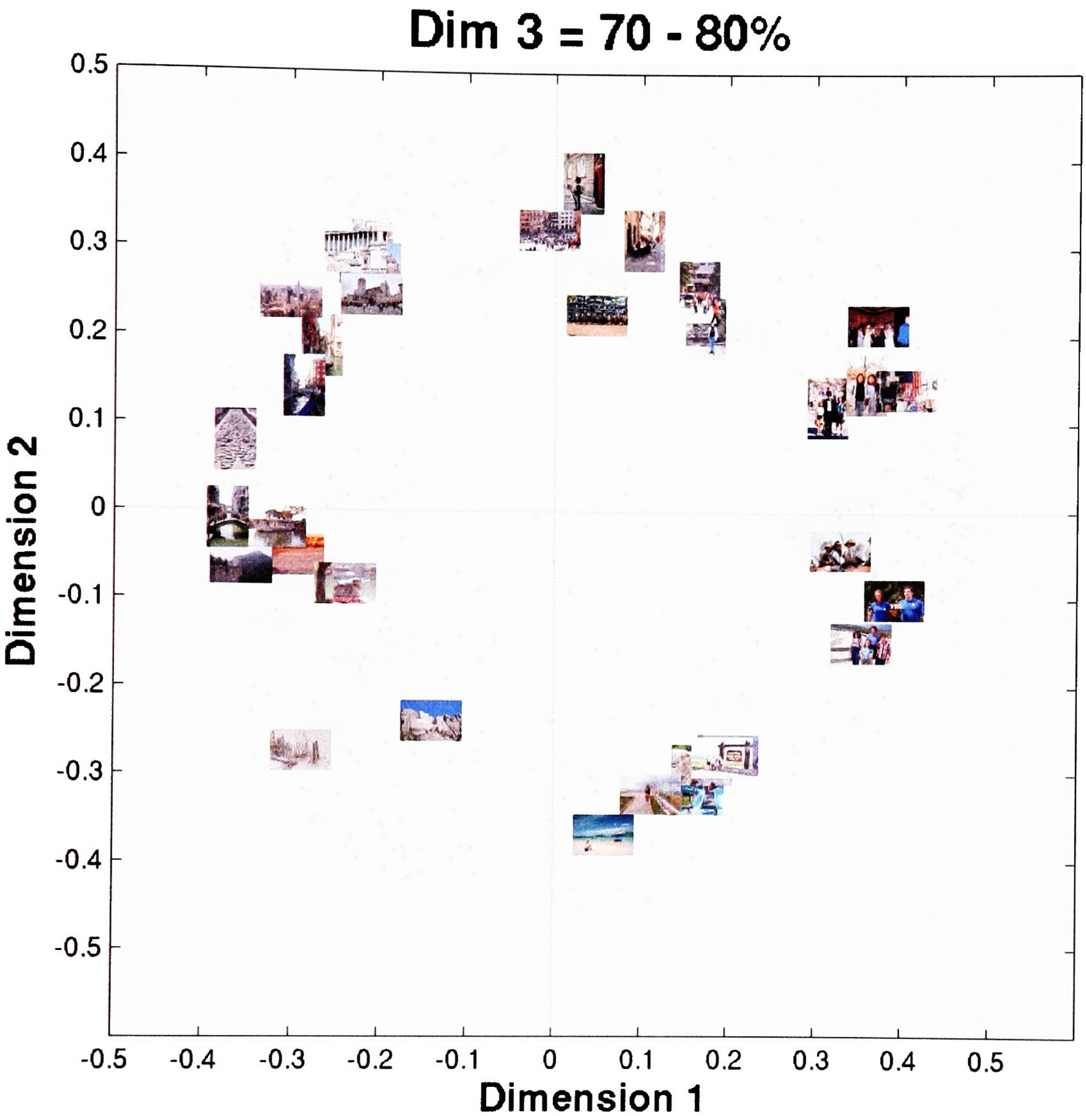


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

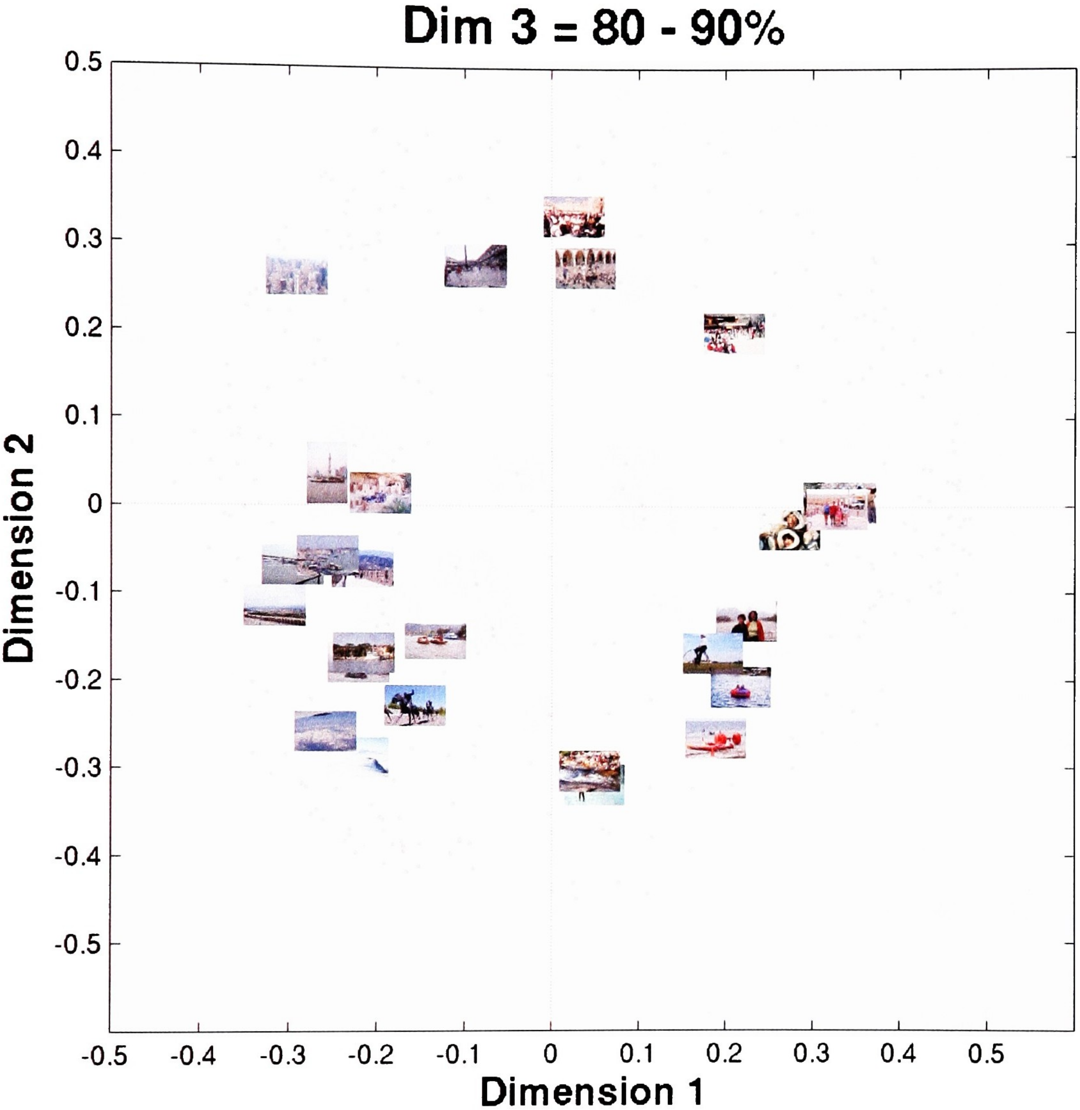


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

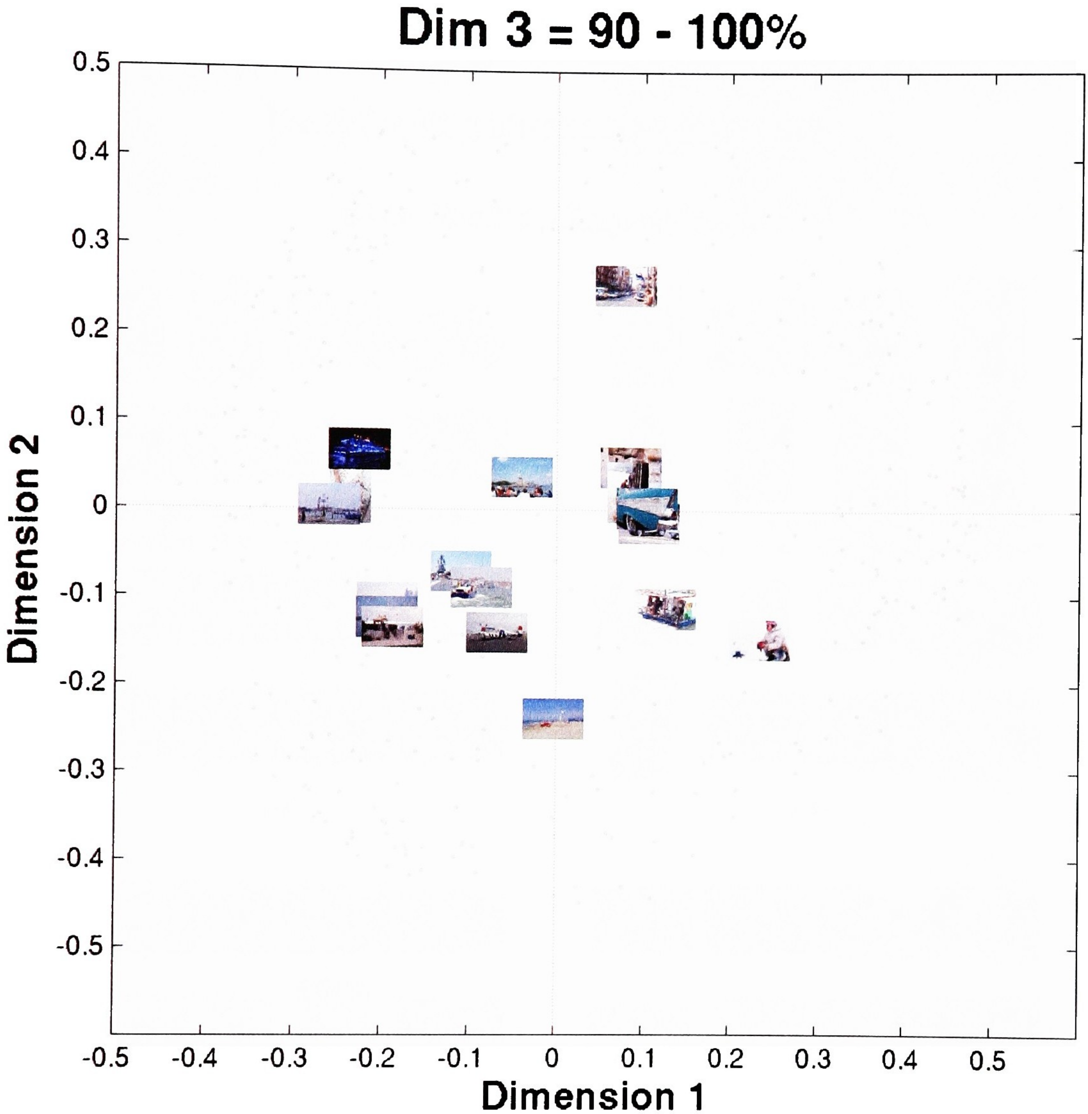


FIGURE D-2 (cont.): Three-dimensional multidimensional scaling configuration for the Distributed Experiment.

APPENDIX E

Dual Scaling Dimension Extremes

Free Sorting Experiment

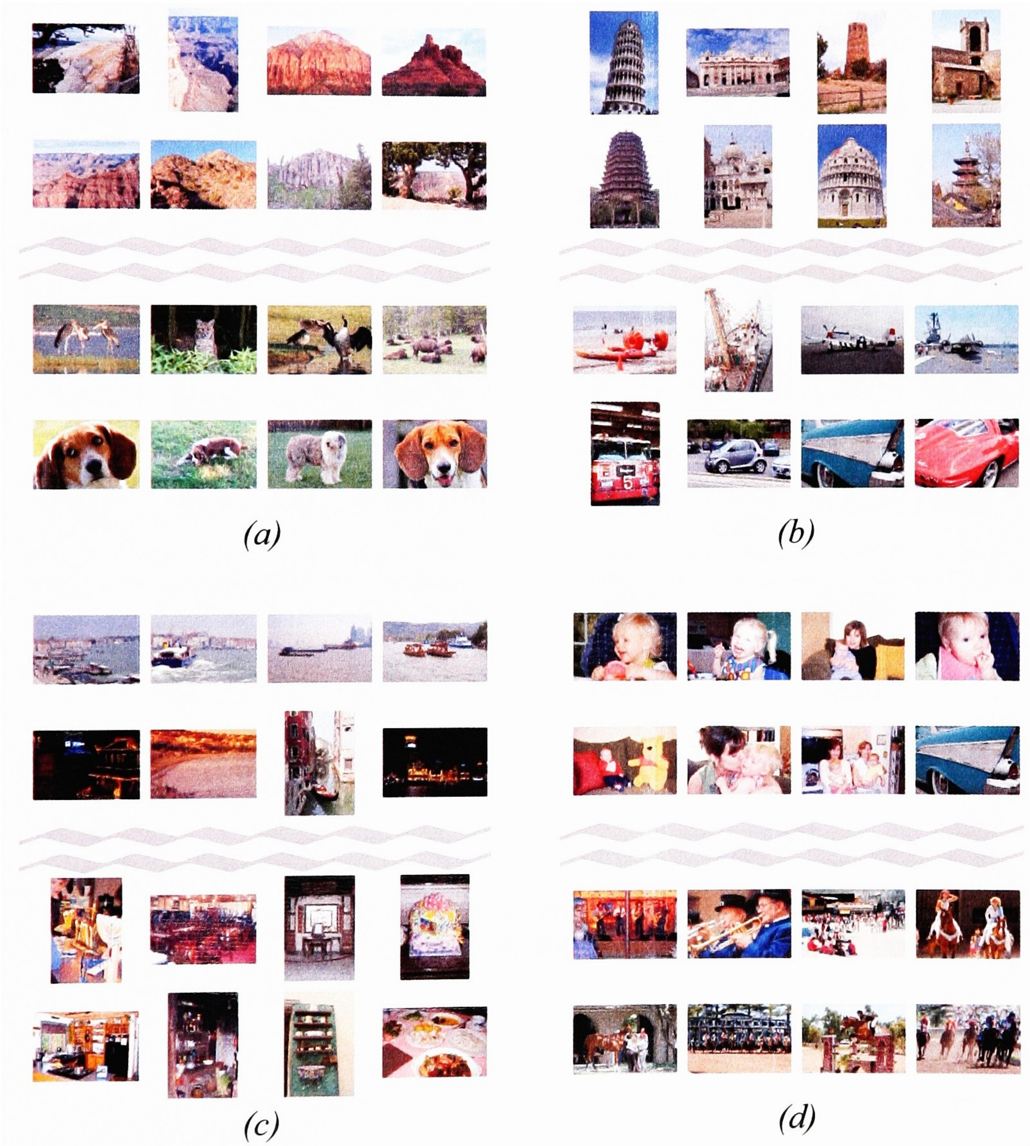


FIGURE E-1 Images at the extremes of (a) dimension 3, (b) dimension 4, (c) dimension 5, and (d) dimension 6 for the Free Sorting Experiment.

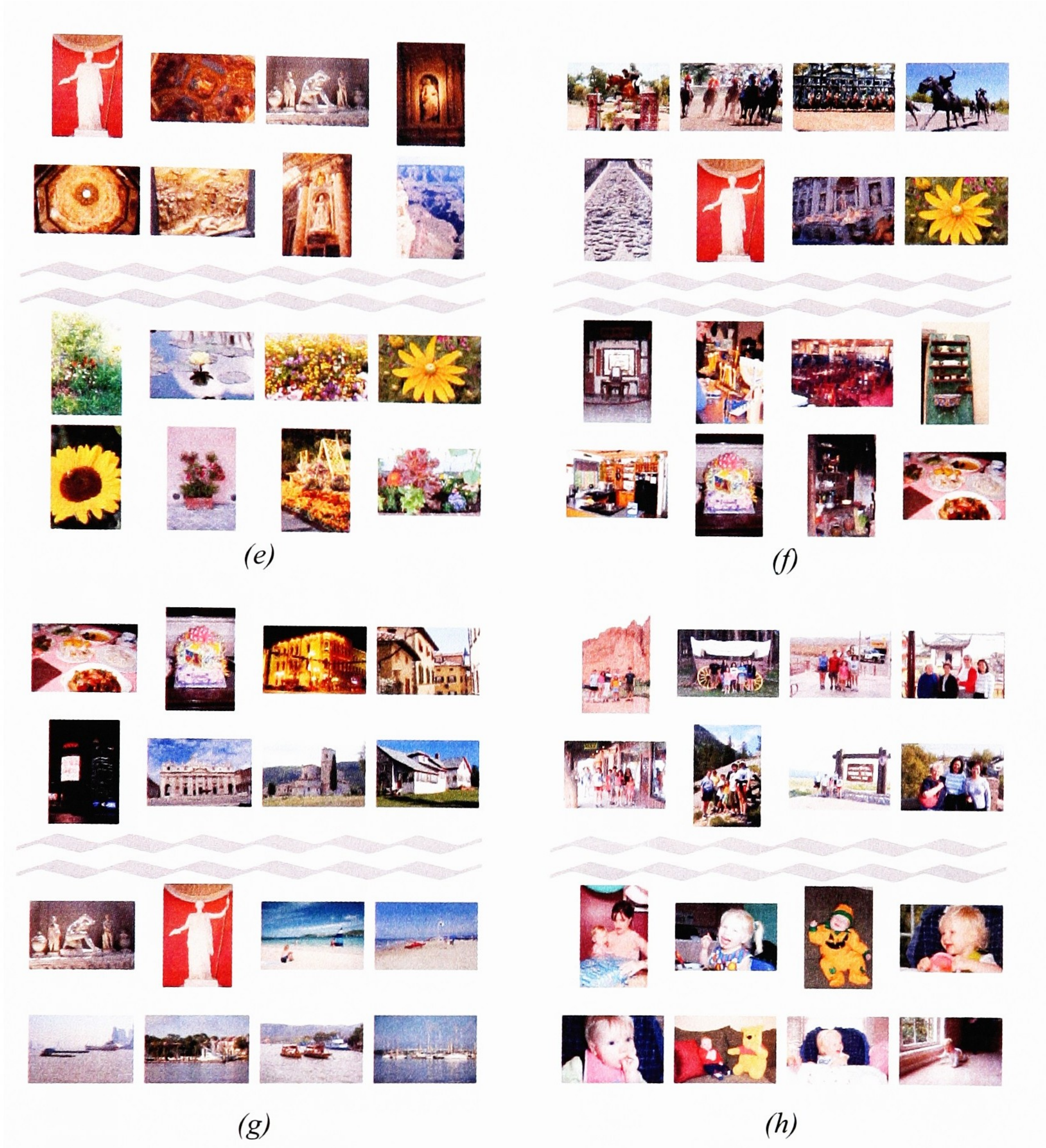


FIGURE E-1 (cont.): Images at the extremes of (e) dimension 7, (f) dimension 8, (g) dimension 9, and (h) dimension 10 for the Free Sorting Experiment.

Distributed Experiment

FIGURE E-2: Images at the extremes of (a) dimension 1, (b) dimension 2, (c) dimension 3, and (d) dimension 4 for the Distributed Experiment.

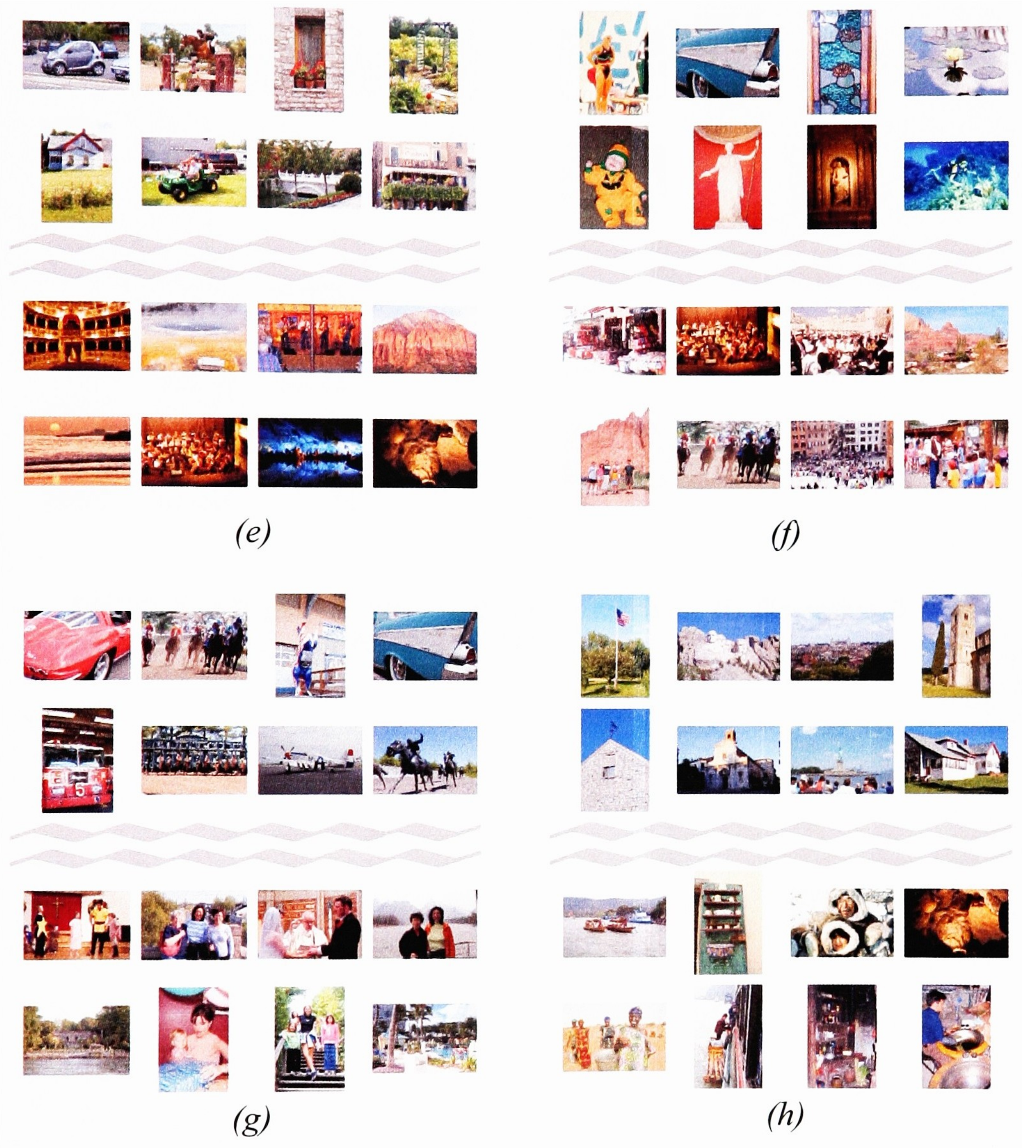


FIGURE E-2 (cont.): Images at the extremes of (e) dimension 5, (f) dimension 6, (g) dimension 7, and (h) dimension 8 for the Distributed Experiment.

FIGURE E-2 (cont.): Images at the extremes of (i) dimension 9 and (j) dimension 6 for the Distributed Experiment.

APPENDIX F

Hierarchical Cluster Analysis — Free Sorting Experiment

5-Cluster Division for the Average Linkage Method

CLUSTER 1

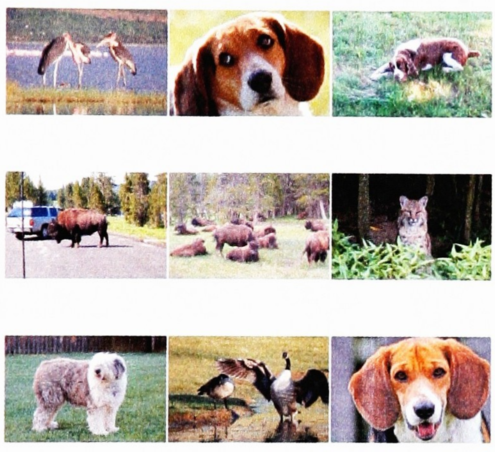


FIGURE F-1: Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 2

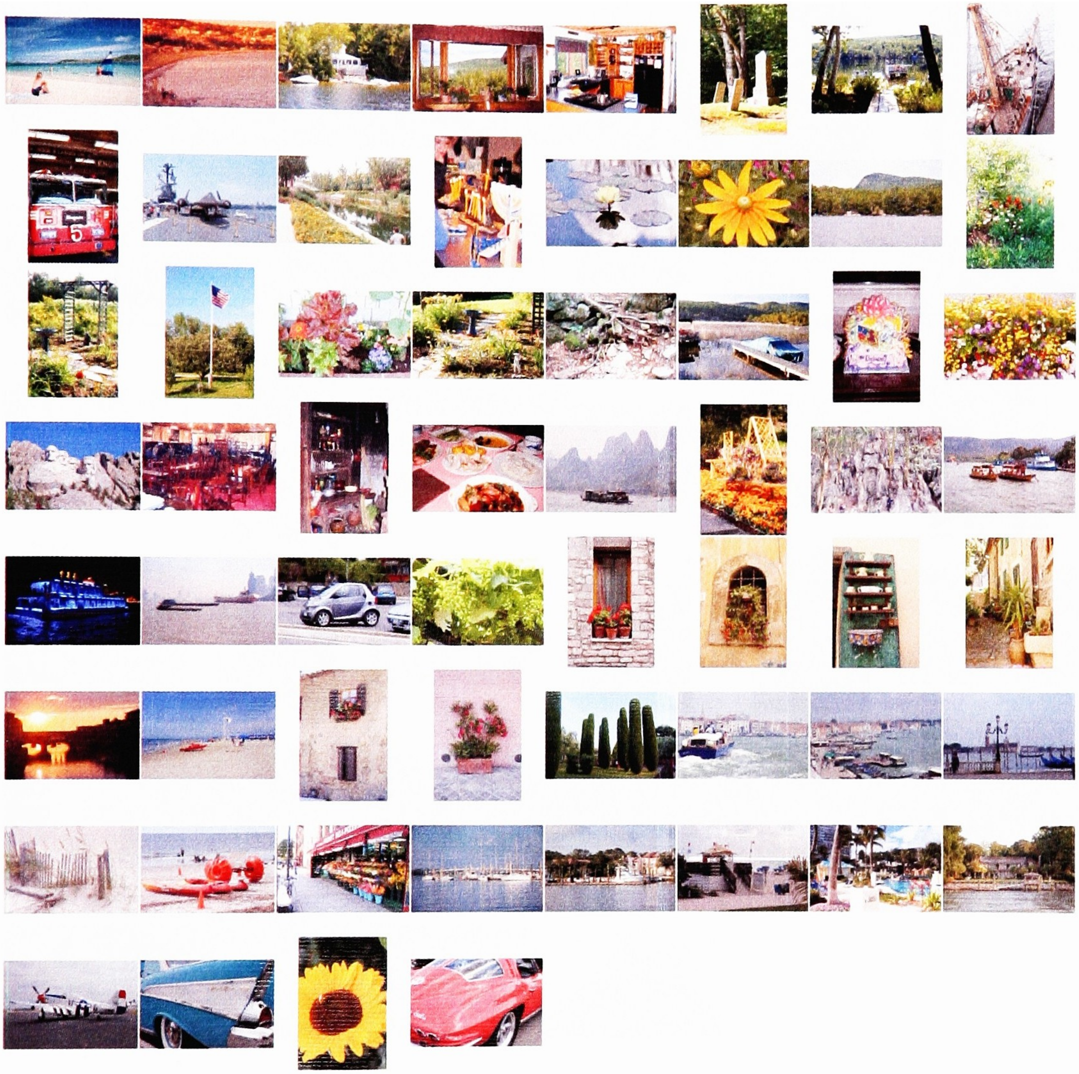


FIGURE F-1 (cont.): Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 3

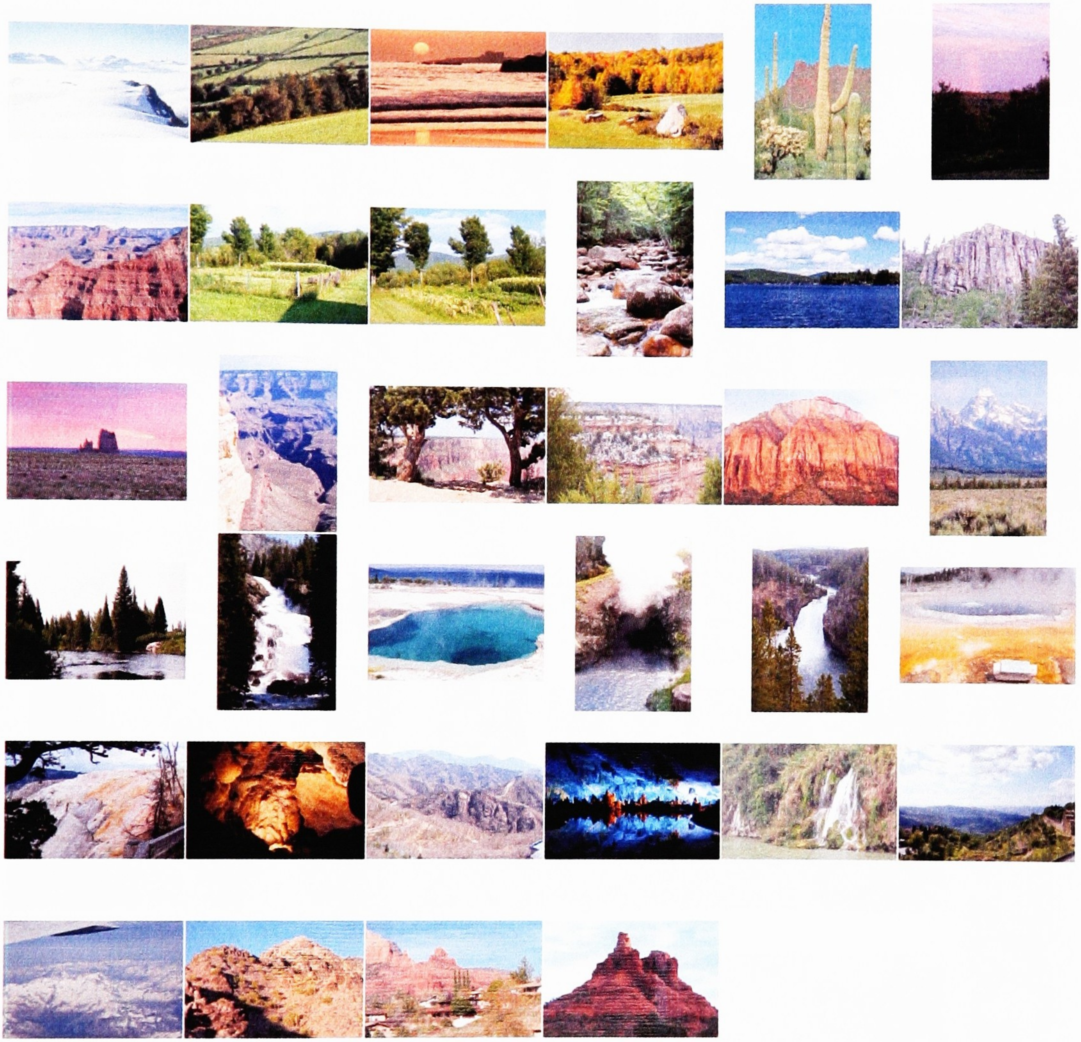


FIGURE F-1 (cont.): Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 4

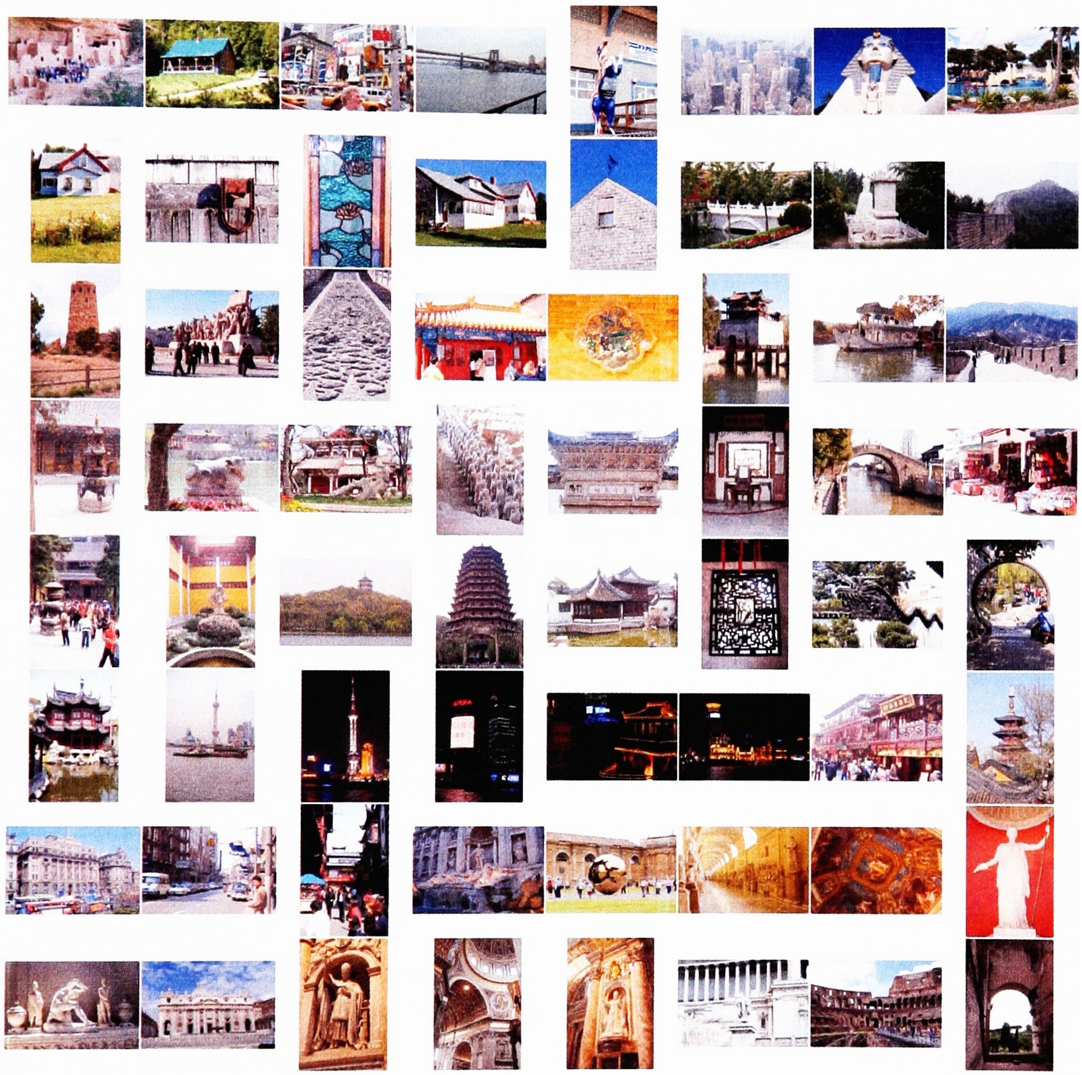


FIGURE F-1 (cont.): Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 4

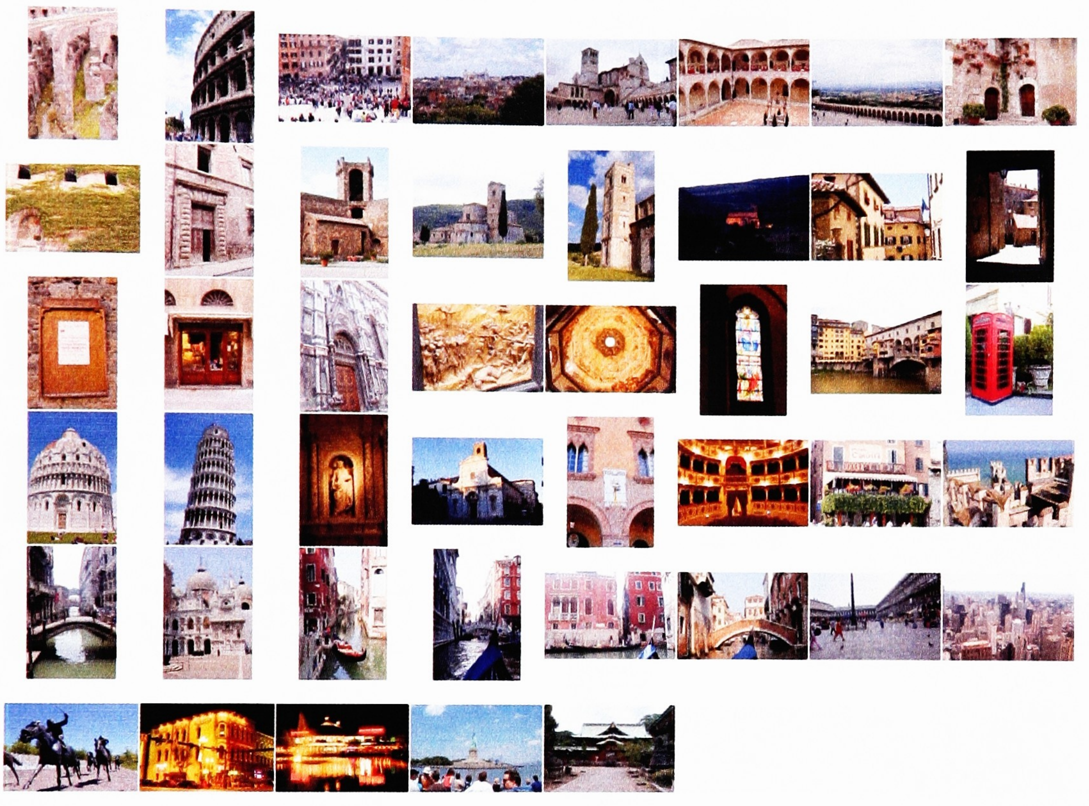


FIGURE F-1 (cont.): Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 5

FIGURE F-1 (cont.): Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 5

FIGURE F-1 (cont.): Hierarchical Cluster Analysis 5-Cluster division using the average linkage method for the Free Sorting Experiment.

6-Cluster Division for the Average Linkage Method

CLUSTER 1

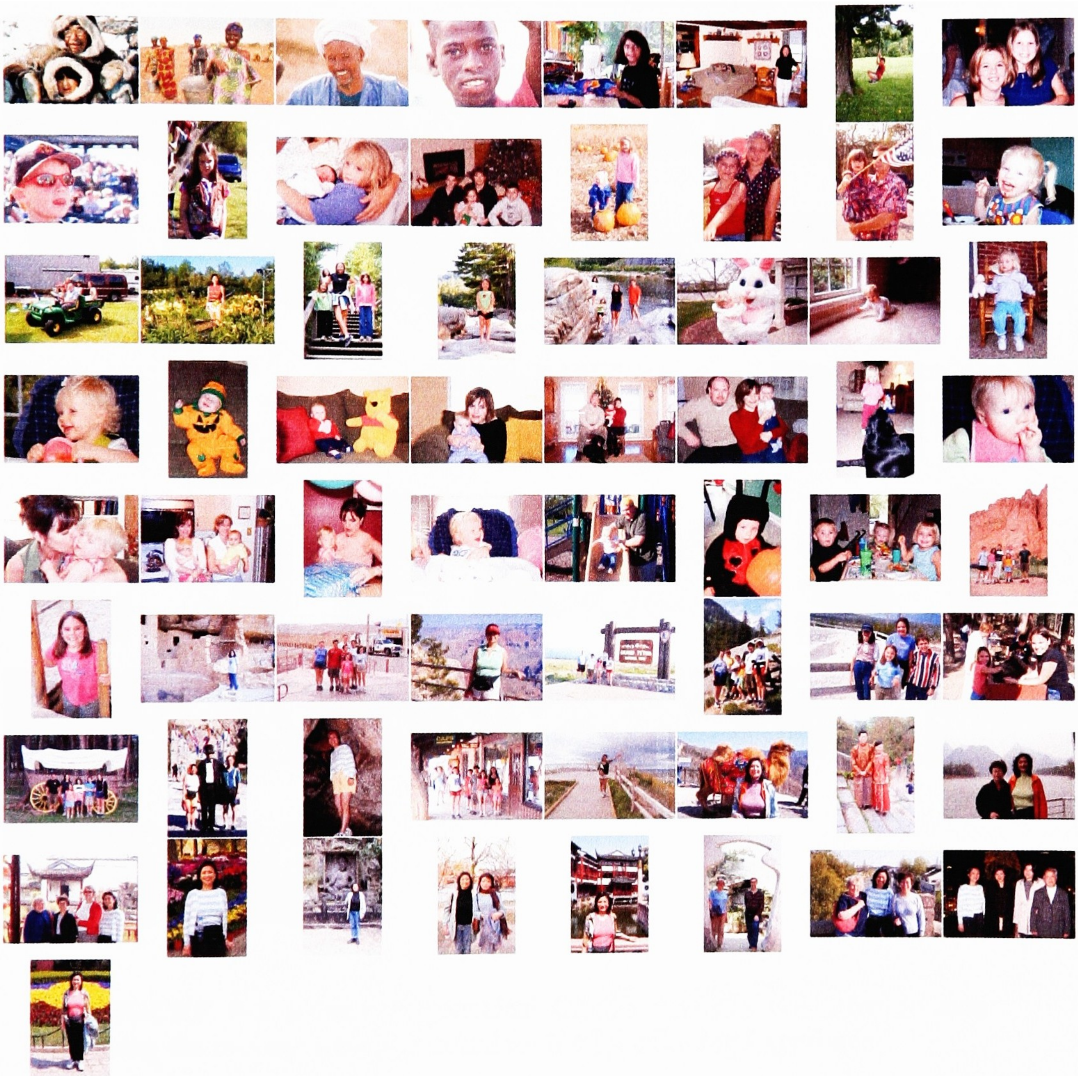


FIGURE F-2: Hierarchical Cluster Analysis 6-Cluster division using the average linkage method for the Free Sorting Experiment.

CLUSTER 2

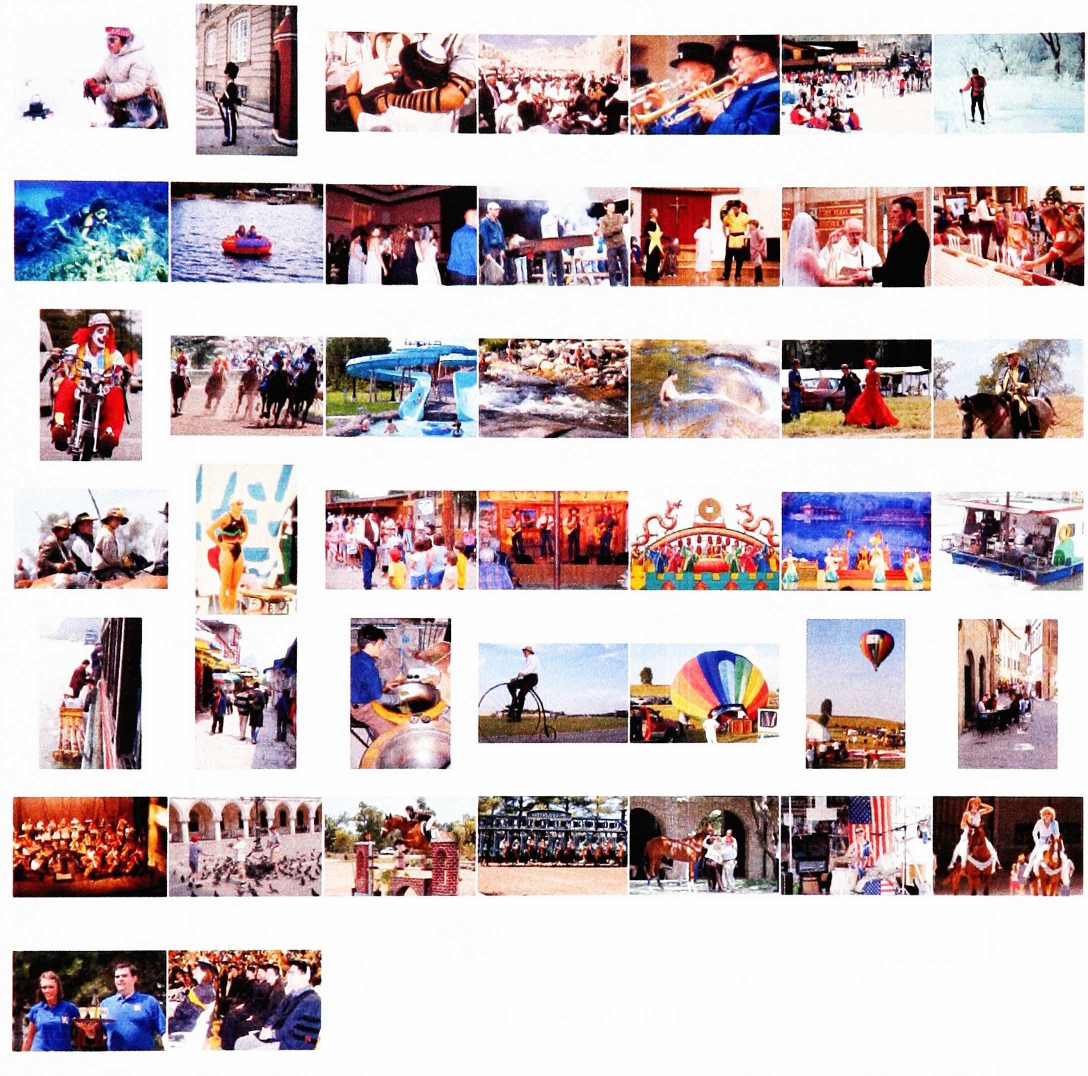


FIGURE F-2 (cont.): Hierarchical Cluster Analysis 6-Cluster division using the average linkage method for the Free Sorting Experiment.

APPENDIX G

Local Multidimensional Scaling — Free Sorting Experiment

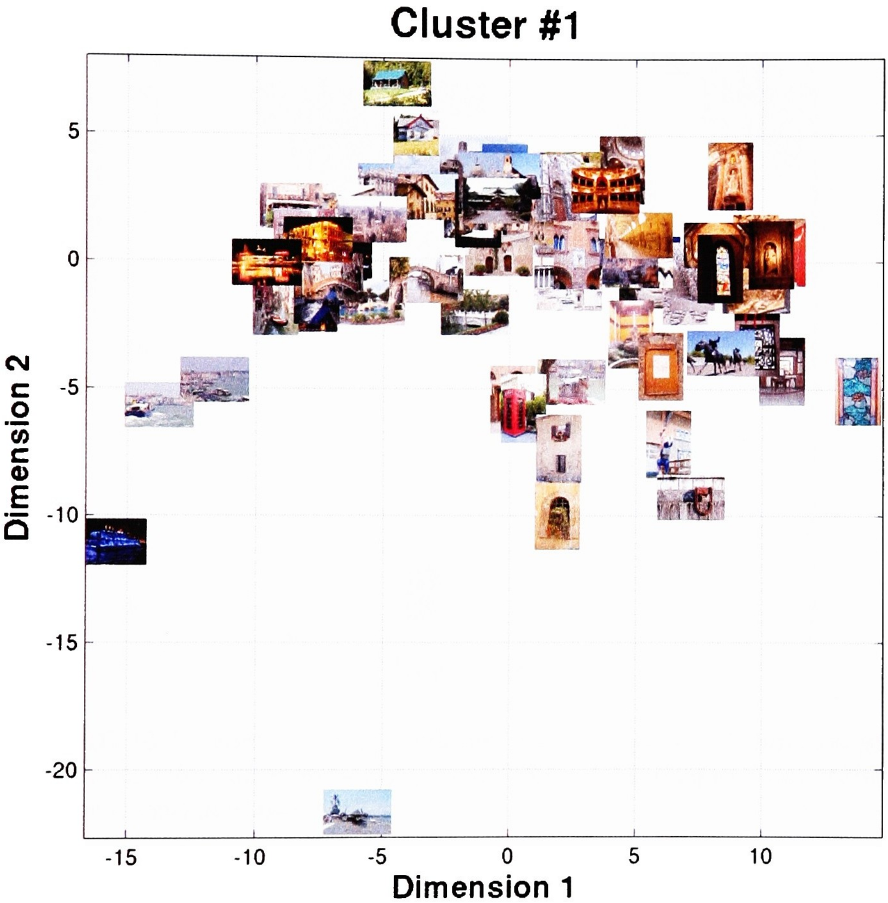


FIGURE G-1: Local multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment using images from within circle #1 (red) in Figure 8-10.

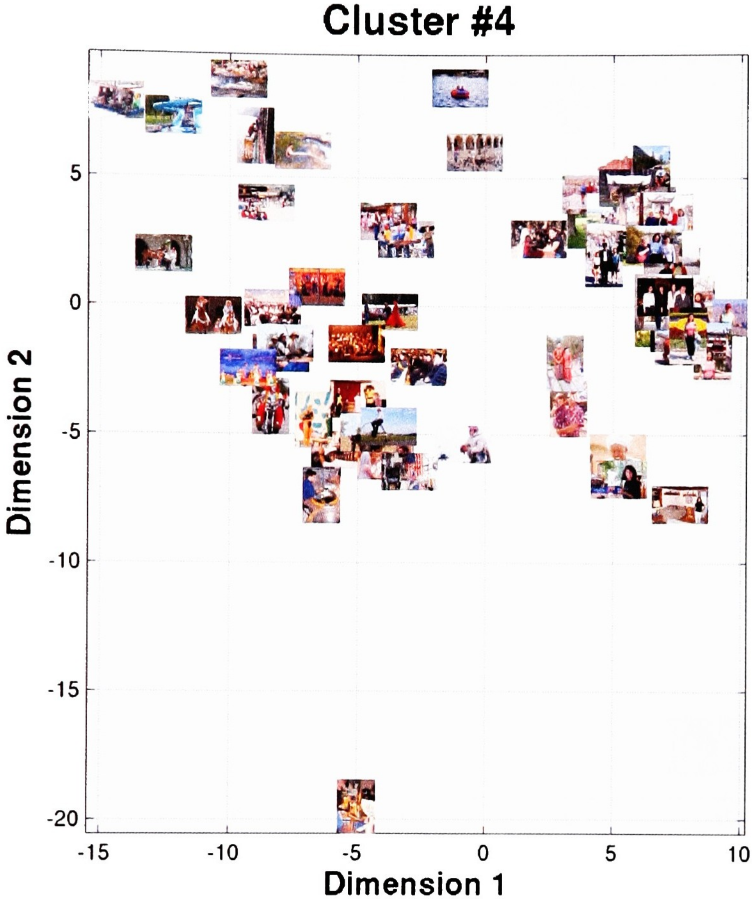


FIGURE G-1 (cont.): Local multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment using images from within circle #4 (cyan) in Figure 8-10.

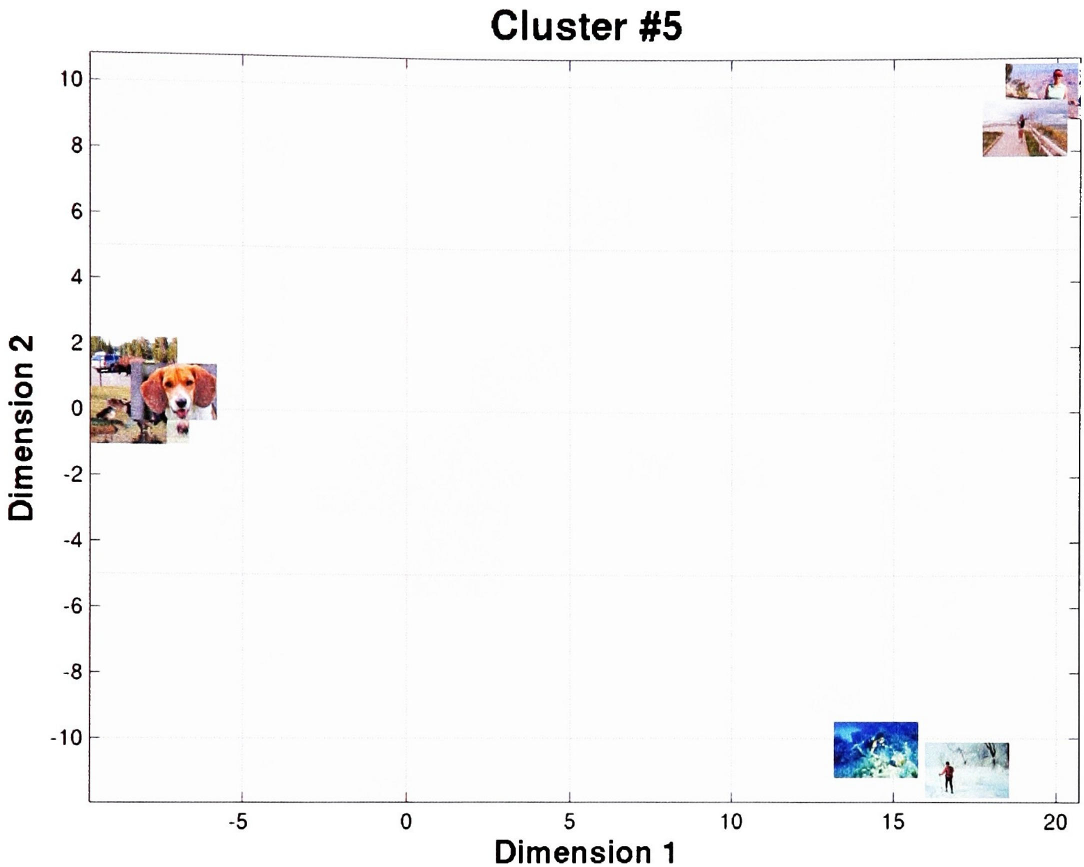


FIGURE G-1 (cont.): Local multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment using images from within circle #5 (magenta) in Figure 8-10.

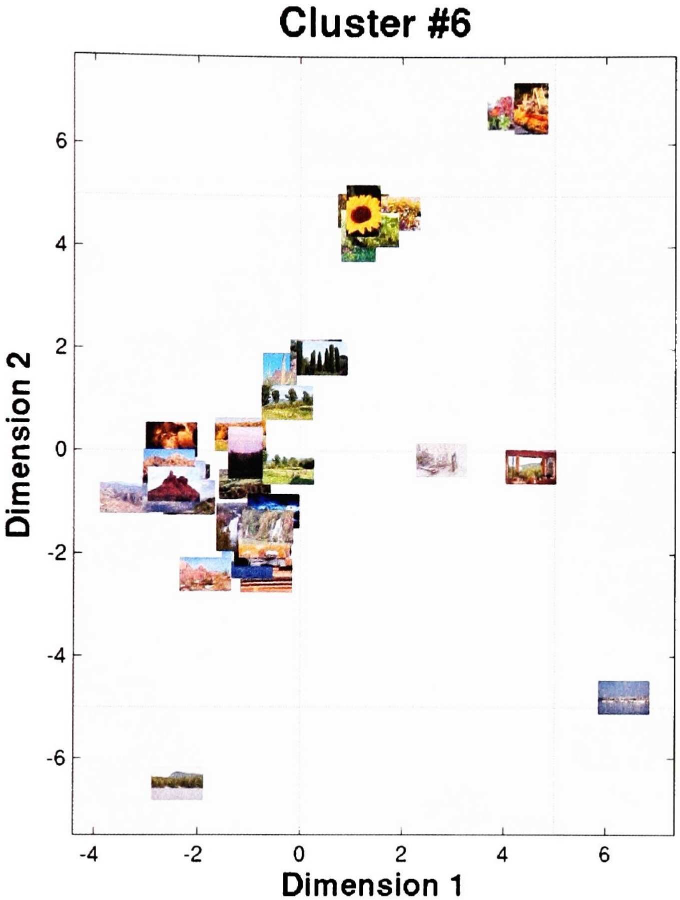


FIGURE G-1 (cont.): Local multidimensional scaling 2-dimensional configuration for the Free Sorting Experiment using images from within circle #6 (black) in Figure 8-10.

Local Multidimensional Scaling — Distributed Experiment

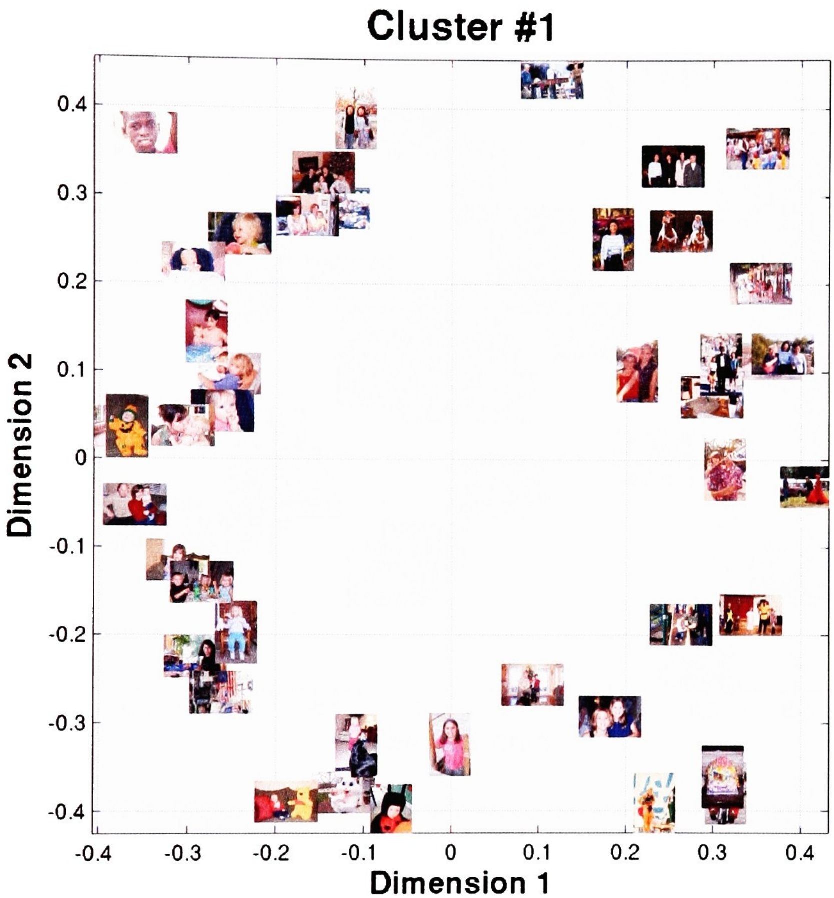


FIGURE G-2: Local multidimensional scaling 2-dimensional configuration for the Distributed Experiment using images from within circle #1 (red) in Figure 8-11.

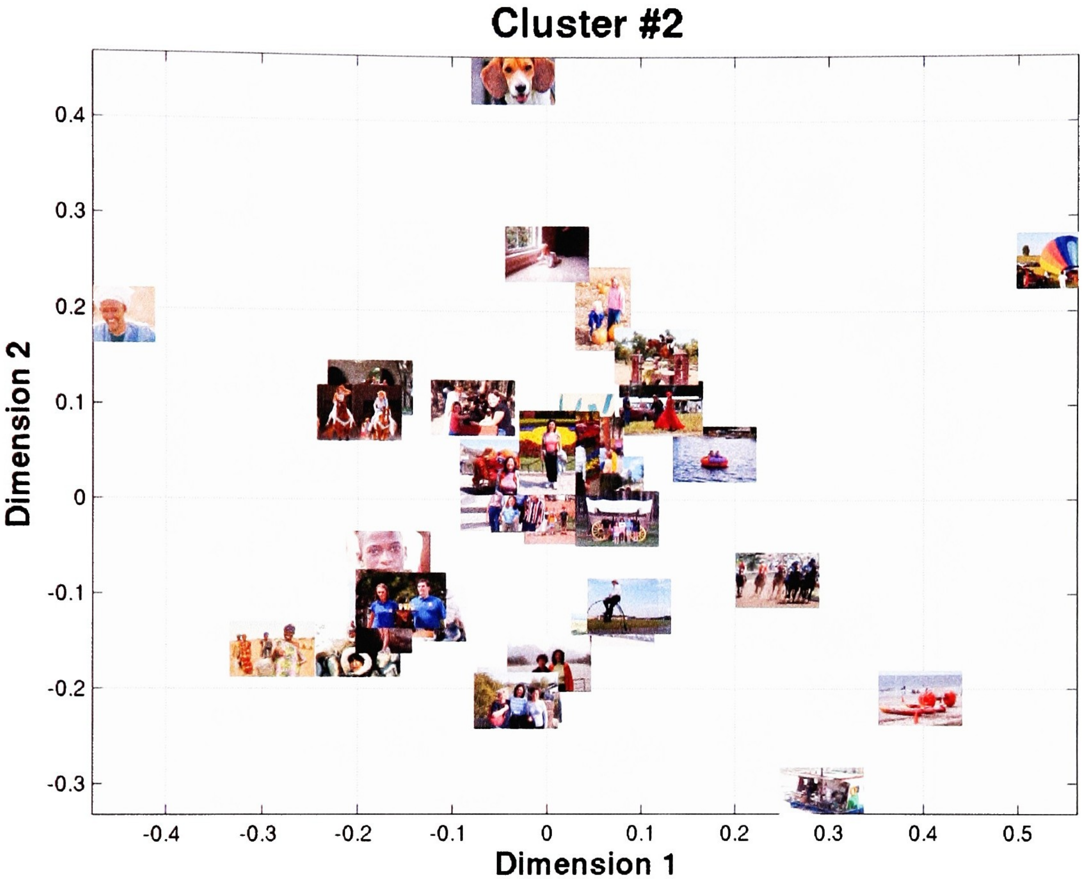


FIGURE G-2 (cont.): Local multidimensional scaling 2-dimensional configuration for the Distributed Experiment using images from within circle #2 (green) in Figure 8-11.

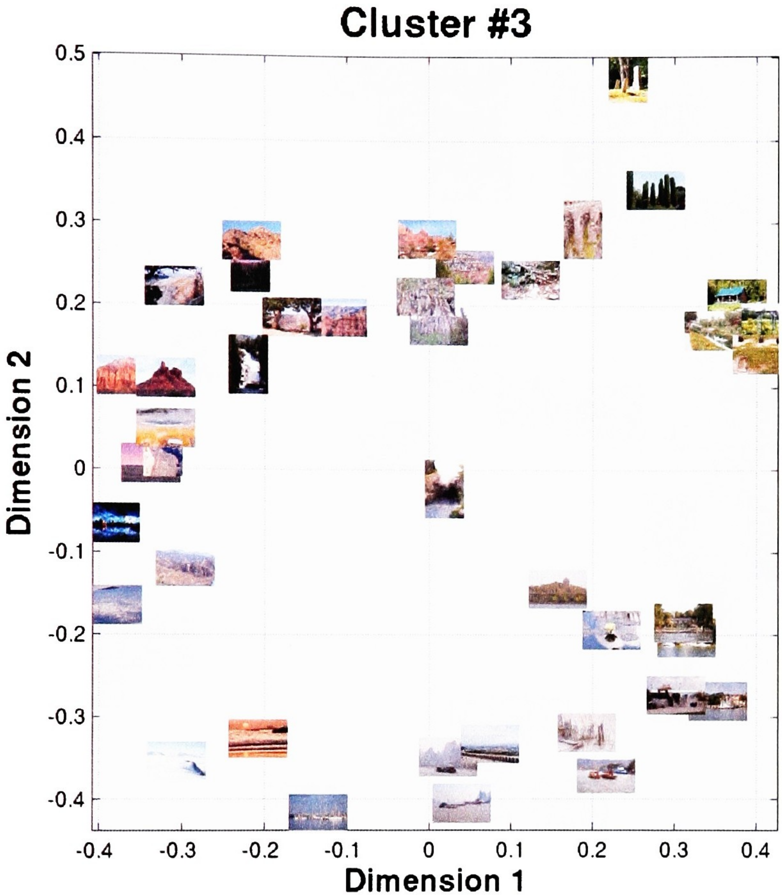


FIGURE G-2 (cont.): Local multidimensional scaling 2-dimensional configuration for the Distributed Experiment using images from within circle #3 (blue) in Figure 8-11.

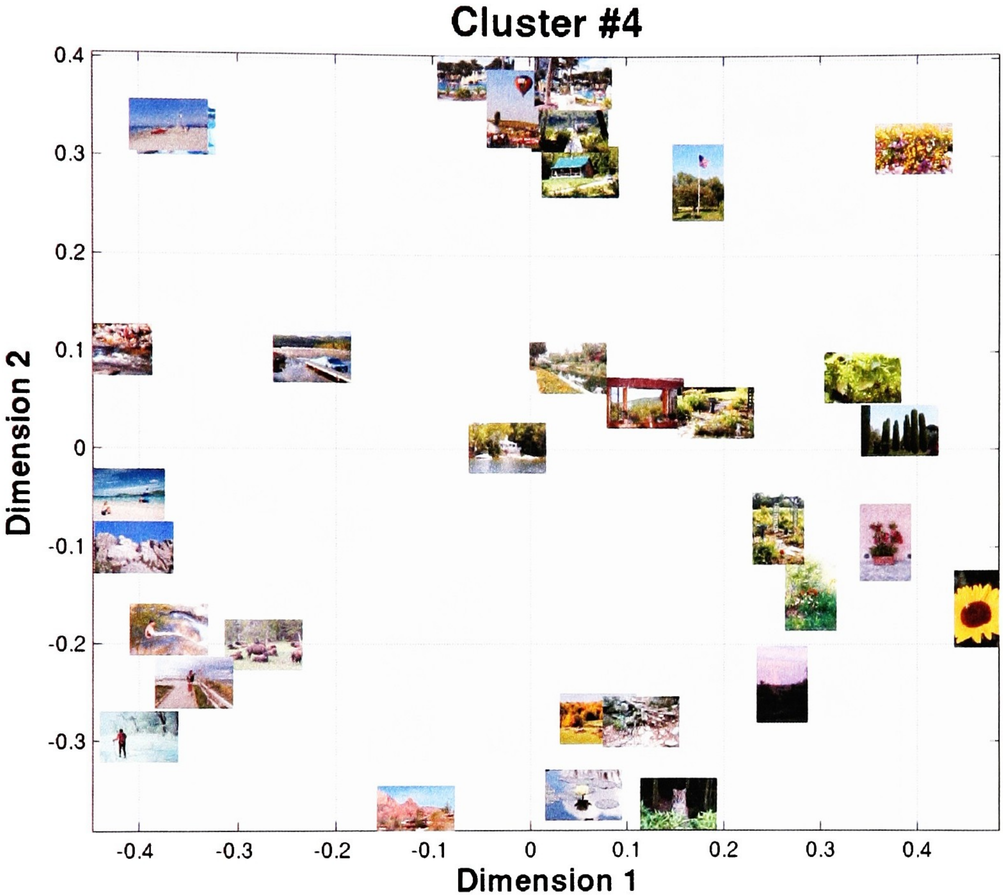


FIGURE G-2 (cont.): Local multidimensional scaling 2-dimensional configuration for the Distributed Experiment using images from within circle #4 (cyan) in Figure 8-11.

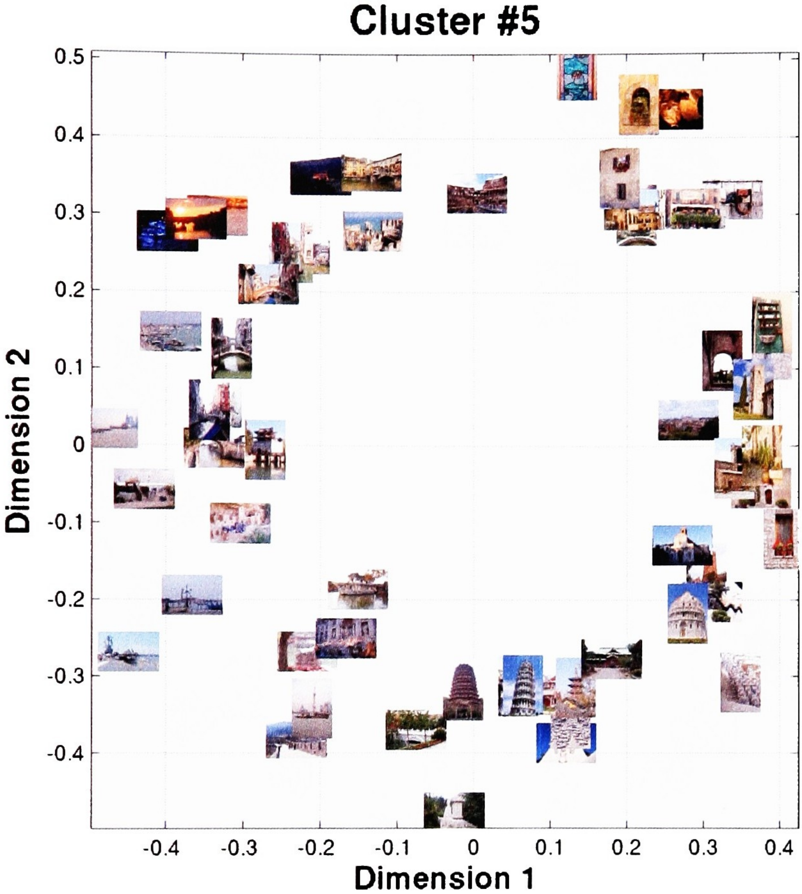


FIGURE G-2 (cont.): Local multidimensional scaling 2-dimensional configuration for the Distributed Experiment using images from within circle #5 (magenta) in Figure 8-11.

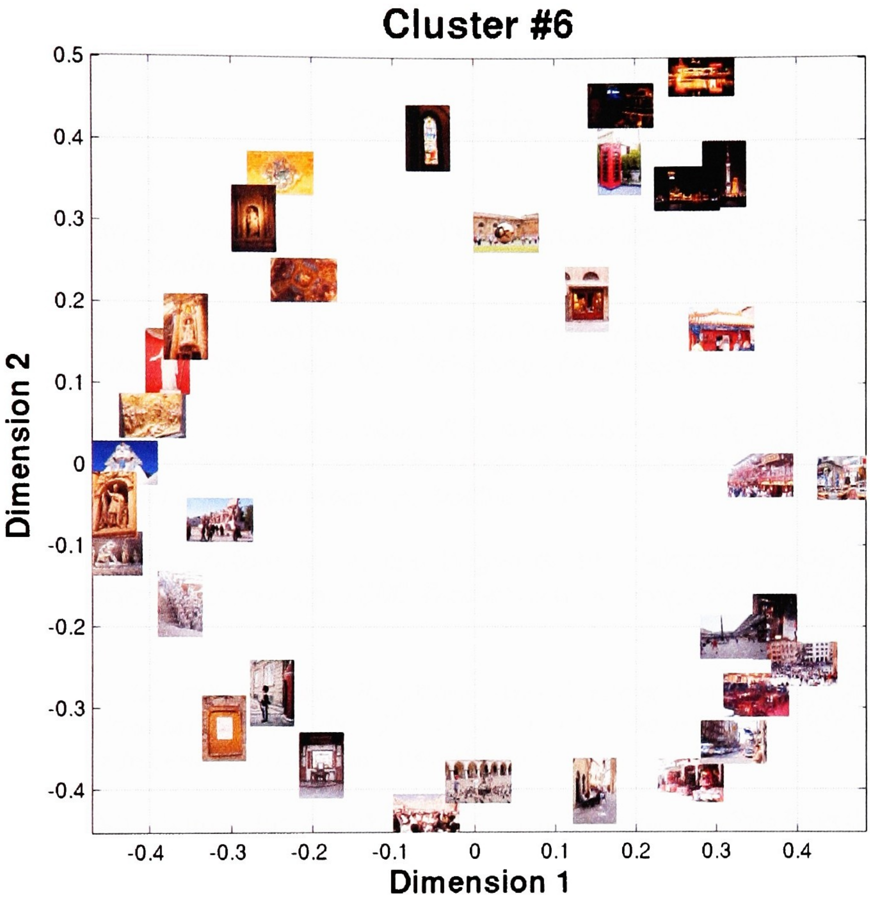


FIGURE G-2 (cont.): Local multidimensional scaling 2-dimensional configuration for the Distributed Experiment using images from within circle #6 (black) in Figure 8-11.

APPENDIX H

References

Berlin, B and Kay, P. *Basic Color Terms; Their Universality and Evolution.* Berkeley: University of California Press; 1969.

Boutell, M, Shen, X, Juo, J, and Brown, C. Multi-Label Semantic Scene Classification, *Technical Report, Dept. Comp. Sci., University of Rochester*; 2003.

Chan, AB, Moreno, PJ, and Vasconcelos, N. Using Statistics to Search and Annotate Pictures: An Evaluation of Semantic Image Annotation and Retrieval on Large Databases, *Joint Statistical Meetings*, Seattle; 2006.

Chen, J, Pappas, TN, Mojsilovic, A, and Rogowitz, BE. Adaptive Perceptual Color-Texture Image Segmentation. *IEEE Transactions on Image Processing*, **14**, 1524-1536; 2005.

Chen, Y, Wang, JZ, and Krovetz, R. Content-Based Image Retrieval by Clustering, MIR '03: *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 193-200; 2003.

Cox, IJ, Miller, ML, Minka, TP, Papathomas, TV, and Yianilos, PN. The Bayesian Image Retrieval System, *PicHunter*: Theory, Implementation, and Psychophysical Experiments, *IEEE Transactions on Image Processing*, **9**, 20-37; 2000.

Depalov, D, Pappas, T, Li, D, and Gandhi, B. Perceptually Based Techniques for Semantic Image Classification and Retrieval, *Proceedings of the SPIE*, **6057**, 354-363; 2006.

Everitt, B. *Cluster Analysis*. London: Heinemann Educational [for] the Social Science Research Council; 1974.

Fairchild, MD. *Color Apearance Models 2nd Edition.* Wiley-IS&T series in imaging science and technology. Chichester, West Sussex, England: J. Wiley; 2005.

Freeman Jr., DH. *Applied Categorical Data Analysis.* New York, NY: Marcel Dekker, Inc.; 1987.

Greisdorf, H and O'Connor, B. What Do Users See? Exploring the Cognitive Nature of Functional Image Retrieval, *Proceedings of the American Society for Information Science and Technology*, **39**, 383-390; 2002.

Iqbal, Q and Aggarwal, JK. Retrieval by Classification of Images Containing Large Manmade Objects Using Perceptual Grouping, *The Journal of the Pattern Recognition Society*, **35**, 1463-1479; 2002.

Johnson, RA and Wichern, DW. *Applied Multivariate Statistical Analysis 5th Edition.* Upper Saddle River, New Jersey: Prentice Hall; 2002.

Kaufman, L and Rousseeuw, PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics. New York: Wiley; 1990.

Kearsley, AJ, Tapia, RA, and Trosset, MW. *The Solution of the Metric STRESS and SSTRESS Problems in Multidimensional Scaling Using Newton's Method.* Ft. Belvoir: Defense Technical Information Center. <http://handle.dtic.mil/100.2/ADA445621>; 1995.

Kruskal, JB. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis, *Psychometrika*, **29**, 1-26; 1964a.

Kruskal, JB. Nonmetric Multidimensional Scaling: A Numerical Method, *Psychometrika*, **29**, 115-129; 1964b.

Kruskal, JB and Wish, M. *Multidimensional Scaling*. Beverly Hills, California: Sage Publications; 1978.

Laine-Hernandez, M and Westman, S. Image Semantics in the Description and Categorization of Journalistic Photographs, *In Grove, A, Eds. Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology*, Austin; 2006.

Le Borgne, H, Guyader, N, Guérin-Dugue, A, and Hérault, J. Classification of Images: ICA Filters vs Human Perception, *Proceedings of 7th International Symposium on Signal Processing and its Applications*, **2**, 251-254; 2003.

Lee, SM, Xin, JH, and Westland, S. Evaluation of Image Similarity by Histogram Intersection. *Color Research and Application*, **30**, 265-274; 2005.

Maranell, GM ed. *Scaling: A Sourcebook for Behavioral Scientists.* Chicago: Aldine Publishing Company; 1974.

Maraun, MD, Slaney, K, and Jalava, J. Dual Scaling for the Analysis of Categorical Data, *Journal of Personality Assessment*, **85**, 209-217; 2005.

Matlab. The Mathworks, Inc., Natick, MA, v. 7.0.0.19901 (R14); 2004.

Mojsilovic, A. A Computational Model for Color Naming and Describing Color Composition of Images, *IEEE Transactions on Image Processing*, **14**, 690-699; 2005.

Mojsilovic, A. A Method for Color Naming and Description of Color Composition in Images, *Proceedings of the International Conference on Image Processing*, **2**, 789-792; 2002.

Mojsilovic, A and Gomes, J. Semantic Based Categorization, Browsing and Retrieval in Medical Image Databases, *Proceedings of the International Conference on Image Processing*, **3**, 145-148; 2002.

Mojsilovic, A, Gomes, J, and Rogowitz, B. ISee: Perceptual Features for Image Library Navigation, *Proceedings of Human Vision and Electronic Imaging VIII*, 266-267; 2002.

Mojsilovic, A, Hu, J. A Method for Color Content Matching of Images, *IEEE International Conference on Multimedia and Expo*, **2**, 649-652; 2000.

Mojsilovic, A, Hu, J, and Soljanin, E. Extraction of Perceptually Important Colors and Similarity Measurement for Image Matching, Retrieval, and Analysis, *IEEE Transactions on Image Processing*, **11**, 1238-1248; 2002.

Mojsilovic, A and Rogowitz, B. Capturing Image Semantics with Low-Level Descriptors, *Proceedings 2001 International Conference on Image Processing*, **1**, 18-21; 2001a.

Mojsilovic, A and Rogowitz, B. A Psychophysical Approach to Modeling Image Semantics, *Proceedings of SPIE*, **4299**, 470-477; 2001b.

Mojsilovic, A and Rogowitz, B. Semantic Metric for Image Library Exploration, *IEEE Transactions on Multimedia*, **6**, 828-838; 2004.

Montag, ED and Kasahara, H. Multidimensional Analysis Reveals Importance of Color for Image Quality, *IS&T/SID 9th Color Imaging Conference*, Scottsdale, 17-21; 2001.

Moroney, N. Unconstrained Web-Based Color Naming Experiment, *Proceedings of the SPIE Color Imaging VIII: Processing, Hardcopy, and Applications*, **5008**, 36-46; 2003.

Moroney, N and Tastl, I. Multi-Dimensional Scaling with Non-Repeating Random Paths, *Proceedings of SPIE-IS&T Electronic Imaging*, **5668**, 20-27; 2005.

Nishisato, S. *Analysis of Categorical Data: Dual Scaling and its Applications*. Mathematical expositions, no. 24. Toronto: University of Toronto Press; 1980.

Nishisato, S. *Elements of Dual Scaling: An Introduction to Practical Data Analysis.* Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1994.

Nishisato, S and Nishisato, I. *Dual Scaling in a Nutshell.* Toronto, Canada: MicroStats; 1994.

Oldfield, D. Kodak "Photo-Space". Personal correspondence; 2005.

Oldfield, DJ, Pino, G, Segur, RK, Odell, S, and Twist, JP. Assessment of the Current Light-Fade Endpoint Metrics Used in the Determination of Print Life – Part I, *Journal of Imaging Science and Technology*, **48**, 495-501; 2004.

Rodden, K. How Do People Organise Their Photographs?, *21st BCS-IRSG Colloquium on IR*, Glasgow; 1999.

Rodden, K, Basalaj, W, Sinclair, D, and Wood, K. Does Organisation by Similarity Assist Image Browsing?, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 190-197; 2001.

Rorissa, A and Hastings, SK. Free Sorting of Images: Attributes Used for Categorization. *Proceedings of the 67th ASIS&T Annual Meeting*, **41**, 360-366; 2004.

Rogowitz, BE, Frese, T, Smith, JR, Bouman, CA, and Kalin, E. Perceptual Image Similarity Experiments, *Proceedings of the SPIE*, **3299**, 576-590; 1998.

Semantics. The Encyclopedia Americana International Edition, **24**, 535; 2001.

Serrano, N, Savakis, A, and Luo, J. A Computationally Efficient Approach to Indoor/Outdoor Scene Classification, *Proceedings of International Conference on Pattern Recognition*, **4**, 146-149; 2002.

Schiffman, SS, Reynolds, ML, and Young, FW. *Introduction to Multidimensional Scaling: Theory, Methods, and Applications.* New York: Academic Press; 1981.

Tian, YH, Huang, TJ, and Gao, W. Exploiting Multi-Context Analysis in Semantic Image Classification, *Journal of Zhejiang University SCIENCE*, **6A**, 1268-1283; 2005.

Torgerson, WS. *Theory and Methods of Scaling.* NewYork: John Wiley & Sons, Inc.; 1958.

Torralba, A and Oliva, A. Statistics of Natural Image Categories, *Network: Computation in Neural Systems*, **14**, 391-412; 2003.

Wardhani, A. Image Searching Tool Using Category-Based Indexing, *Proceedings of APRS Workshop on Digital Image Computing*, 99-104; 2003.

Wardhani, A and Thomson, T. Content Based Image Retrieval Using Category-Based Indexing, *IEEE International Conference on Multimedia and Expo*, **2**, 783-786; 2004.

Young, FW and Hamer, RM. *Multidimensional Scaling: History, Theory and Applications*. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1987.